Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity*

Bruno Ferman†
Sao Paulo School of Economics - FGV

Cristine Pinto‡
Sao Paulo School of Economics - FGV

First Draft: October, 2015
This Draft: December, 2015

Abstract

Differences-in-Differences (DID) is one of the most widely used identification strategies in applied economics. However, inference in DID models when there are few treated groups remains an open question. We show that the usual inference methods used in DID models might not perform well when there are few treated groups and errors are heteroskedastic. In particular, we show that when there is variation in the number of observations per group, inference methods designed to work when there are few treated groups tend to (under-) over-reject the null hypothesis when the treated groups are (large) small relative to the control groups. This happens because larger groups tend to have lower variance, generating heteroskedasticity in the group x time aggregate DID model. We provide evidence from Monte Carlo simulations and from placebo DID regressions with the American Community Survey (ACS) and the Current Population Survey (CPS) datasets to show that this problem is relevant even in datasets with large numbers of observations per group. We then derive an alternative inference method that provides accurate hypothesis testing in situations where there are few treated groups (or even just one) and many control groups in the presence of heteroskedasticity. Our method assumes that we know how the heteroskedasticity is generated, which is the case when it is generated by variation in the number of observations per group. We only need to know the structure of the heteroskedasticity of a linear combination of the errors, which implies that we do not need strong assumptions on the intra-group and serial correlation structure of the errors. Our method provided accurate hypothesis testing with one treated and 24 control groups in simulations with real datasets. Finally, we also show that an inference method for the Synthetic Control Estimator proposed by Abadie et al. (2010) can correct for the heteroskedasticity problem, and derive conditions under which this inference method provides accurate hypothesis testing.

Keywords: differences-in-differences; inference; heteroskedasticity; clustering; few clusters; bootstrap; synthetic control

JEL Codes: C12; C21; C33

*We would like to thank Josh Angrist, Sergio Firpo, Bernardo Guimaraes, Lance Lochner, Vladimir Ponczek, Andre Portela, Vitor Possebom, Rodrigo Soares, Chris Taber, Gabriel Ulyssea and seminar participants at Sao Paulo School of Economics - FGV and PUC-Rio for comments and suggestions.

†bruno.ferman@fgv.br

‡cristine.pinto@fgv.br

1 Introduction

Differences-in-Differences (DID) is one of the most widely used identification strategies in applied economics. However, inference in DID models is complicated by the fact that residuals might exhibit intra-group and serial correlations. Not taking these problems into account can lead to severe underestimation of the DID standard errors, as highlighted in Bertrand et al. (2004). Still, there is as yet no unified approach to deal with this problem. As stated in Angrist and Pischke (2009), *"... there are a number of ways to do this [deal with the serial correlation problem], not all equally effective in all situations. It seems fair to say that the question of how best to approach the serial correlation problem is currently under study, and a consensus has not yet emerged."*

One of the most common solutions to this problem is to use the cluster-robust variance estimator (CRVE) at the group level.[1] By clustering at the group level, we allow for unrestricted correlation in the within-group errors. More specifically, we allow not only for correlation in the errors of observations in the same group x time, but also for correlation in errors of observations in the same group at different time periods. One important advantage of CRVE is that it also allows for unrestricted heteroskedasticity. The variance of the DID estimator can be divided into two components: one related to the variance of the treated groups and another one related to the variance of the control groups. The CRVE takes heteroskedasticity into account by essentially estimating the variance separately for the treated and for the control groups. Bertrand et al. (2004) show that CRVE and pairs-bootstrap at the group level work well when the number of groups is large.[2] Even when there are only a small number of groups, it might still be possible to obtain tests with correct size even with unrestricted heteroskedasticity (Cameron et al. (2008), Brewer et al. (2013), Imbens and Kolesar (2012), Bell and McCaffrey (2002), Canay et al. (2014), and Ibragimov and Mller (2013)). However, these inference methods will eventually fail when the proportion of treated groups goes to zero or one, even if there are many groups in total (MacKinnon and Webb (2015b)). The problem is that, with a small number of treated groups, it is hard to estimate the variance component related to the treated groups based only on the residuals of the treated group. In the polar case where there is only one treated group, the CRVE estimate of this component of the variance would be identical to zero.[3]

[1] The CRVE was developed by Liang and Zeger (1986). We can think of this method as a generalization of the heterocedasticity-robust variance matrix due to White (1980). In typical applications the label "group" stands for states, counties or countries. More generally, we refer to group as the unit level that is treated. We will assume throughout that errors of individuals within a group can be correlated while errors of individuals in different groups are uncorrelated.

[2] Wooldridge (2003) provides an overview of cluster-sample methods in linear models. The author shows that when the number of groups increases and the groups sizes are fixed, the theory is well developed.

[3] Another alternative presented by Bertrand et al. (2004) is to collapse the pre- and post-information. This approach would take care of the auto-correlation problem. However, in order to allow for heteroskedasticity, one would have to use robust standard errors, in which case this method would also fail when there are few treated groups.

An alternative when there are few treated groups is to use information from the control groups in order to estimate the component of the variance related to the treated groups. Donald and Lang (2007), henceforth DL, deal with the case when the number of treated and control groups is small. They use small sample inference procedures on the group x time DID aggregate model under the assumption that errors are normal, homoskedastic and serially uncorrelated. Conley and Taber (2011), henceforth CT, provide an interesting inference method to take both intra-group and serial correlations into account when the number of treated groups is small, but the number of control groups is large. Their method uses information on the residuals of the control groups to estimate the distribution of the DID estimator under the null. Cluster residual bootstrap provides another alternative when there are few treated clusters (Cameron et al. (2008)). In cluster residual bootstrap, we hold the treatment variable constant throughout the pseudo-samples, while resampling the residuals, so that we guarantee that every pseudo-sample will have the same number of treated groups. A crucial assumption for all these methods is that the errors (or a linear combination of the errors) are homoskedastic, so that we can use information on the residuals of the control group to assess the variance of the treated group. However, this homoskedasticity assumption might be very restrictive in DID applications. In particular, errors in the group x time DID aggregate model should be inherently heteroskedastic when there is variation in the numbers of observations used to calculate each group x time average.

In this paper, we first show that usual inference methods used in DID models might not perform well when the number of treated groups is small. Methods that allow for unrestricted heteroskedasticity do not work because they estimate the component of the variance related to the treated groups based on few observations. Also, alternative methods that use information from the control groups will not work properly in the presence of heteroskedasticity. In the particular case in which the number of observations per group varies, these methods tend to (under-) over-reject the null hypothesis when the number of observations in the treated groups is (large) small relative to the number of observations in the control groups. The problem is that variation in the number of observations per group invalidates the homoskedasticity assumption, because larger groups tend to have lower variance. The intuition of this result was already articulated in Assuncao and Ferman (2015) in an application of CT.[4] We formalize this idea and derive conditions under which this

[4]Assuncao and Ferman (2015) exclude the comparison of placebo estimates when the placebo treated group is much smaller than the original treated group. As stated in Assuncao and Ferman (2015), "*One important caveat with this method [Conley and Taber (2011)] is that the number of observations in each treatment group × year cell in the placebo regressions will not be the same as in the original regression. This is particularly important when the number of observations in the treatment group is small relative to the control group. In this case, increasing the number of observations in the treatment group would reduce the variance of the estimator even if we hold the number of observations constant. If this correction is not used, then a placebo estimator using a state with few observations as the treatment group would have a much higher variance than our actual*

problem would be more or less relevant. In particular, we show that this problem becomes more severe when the intra-group correlation is smaller and when there are fewer observations per group. We then provide evidence from Monte Carlo simulations and simulations with real datasets to show that this problem can be relevant even in datasets with very large numbers of observations per group. This occurs because, as the intra-group correlation approaches zero, increasing the number of observations per group has little impact on the heteroskedasticity. Therefore, a large number of individual observations per group should not be a reasonable justification for the assumption that group x time averages have homoskedastic residuals.

We then derive an alternative method for inference when there are only few treated groups that takes into account the fact that errors are inherently heteroskedastic when there is variation in the number of observations per group (including the case of only one treated group). The main assumption is that we know how the heteroskedasticity is generated, which is the case when it is generated by variation in the number of observations per group. Under this assumption, we can re-scale the residuals of the control groups using the (estimated) heteroskedasticity structure in a way that allows us to use this information to estimate the distribution of the error for the treated groups. Our method only requires information on the heteroskedasticity structure for a linear combination of the errors, which implies that we do not have to impose strong assumptions on the serial correlation structure. Therefore, our method is more robust than econometric corrections that place a specific parametric form on the time-series process either to estimate the standard errors or to run a FGLS.[5] We show that a cluster residual bootstrap with this heteroskedasticity correction provides valid hypothesis testing asymptotically when the number of control groups goes to infinity. Our Monte Carlo simulations and simulations with real datasets suggest that our method provides hypothesis testing with correct sizes when there are around 25 groups in total (1 treated and 24 controls). We also show that the power of our test converges to the power of the uniformly most powerful test (UMP) when the number of control groups increases.

Our method is closely related to the Randomization Inference (RI) approach proposed by Fisher (1935). In this approach, one uses a permutation test that calculates the test statistic under all possible combinations of treatment assignment, and rejects the null if the observed realization in the actual experiment is extreme

estimator, while a placebo estimator using a large state as the treatment group would tend to underestimate this variance."

[5] Bertrand et al. (2004) show that parametric corrections do not perform well because the coefficient on the auto-correlation parameter is downward biased and because the time-series process might not be correctly specified. Hansen (2007) proposes a bias correction for the auto-correlation estimators. Hausman and Kuersteiner (2008) use a second order expansion to provide a FGLS t-test that takes into consideration the fact that the covariance matrix of the errors has been estimated. Brewer et al. (2013) show that FGLS with Hansen (2007) bias correction combined with robust inference can produce tests with correct sizes even with few groups. However, their approach relies on using the FGLS residuals into the CRVE formula. Therefore, their method would not be appropriate when the proportion of treated groups goes to zero or one, which is the case analyzed in this paper.

enough. The RI approach assumes that treatment assignment is the only stochastic element of the model. In this case, RI provides exact hypothesis testing regardless of the characteristics of the residuals (Lehmann and Romano (2008)). The RI test remains valid in the presence of heteroskedasticity for unconditional tests (that is, before we know which groups were treated). However, once one has information on the size of the treated groups, one should then incorporate this information into the test, as argued in Yates (1984). More specifically, given that one knows that the treated groups are (large) small, one would have information that a permutation test that does not take this information into account would (under-) over-reject the null when the null is true. Canay et al. (2014) show that it would be possible to incorporate this information if one had functions of the data that have the same limiting distribution under the null hypothesis in all permutations. However, we argue that alternative permutation methods that incorporate this information are not feasible when there are very few treated groups. In a recent paper, MacKinnon and Webb (2015a) suggest a permutation test on a t-statistic, which is constructed using CRVE. Their method works when the numbers of treated and control groups are large enough, as asymptotically the t-statistic have the same distribution under the null for all permutations. However, their method does not work well when there are only very few treated groups.[6] The key point is that we go back to the original problem of estimating the variance of the treated groups using CRVE with few treated groups. In contrast, our method provides a valid correction for heteroskedasticity even when there is only one treated group.

Finally, we show that Synthetic Control, an alternative estimation method for the case of one treated group proposed by Abadie et al. (2010), can provide accurate hypothesis testing even in presence of heteroskedasticity. This happens because, under some circumstances, an inference method proposed in Abadie et al. (2010) turns out to correct for the presence of heteroskedasticity by using information from the pre-treatment period. We derive the conditions under which this method provides accurate hypothesis testing. One important scenario that Abadie et al. (2010) does not correct for heteroskedasticity (and our method does) is when there is only one pre-treatment period.

The remainder of this paper proceeds as follows. In Section 2 we present our base model. We briefly explain the necessary assumptions in the existing inference methods, and explain why heteroskedasticity usually invalidates inference methods designed to deal with the case of few treated groups. Then we derive an alternative inference method that corrects for heteroskedasticity even when there is only one treated group. We also derive the conditions under which the inference method for Synthetic Control proposed by

[6] In particular, MacKinnon and Webb (2015a) method is essentially the same as CT method when there is only one treated group.

Abadie et al. (2010) provides accurate hypothesis testing in the presence of heteroskedasticity. In Section 3 we perform Monte Carlo simulations to examine the performance of existing inference methods and to compare that to the performance of our method with heteroskedasticity correction. In Section 4 we compare the different inference methods by simulating placebo laws in real datasets: the American Community Survey (ACS) and the Current Population Survey (CPS). We conclude in Section 5.

2 Base Model

2.1 A Review of Existing Methods

We consider a group x time DID aggregate model:[7]

$$Y_{jt} = \alpha d_{jt} + \theta_j + \gamma_t + \eta_{jt} \tag{1}$$

where Y_{jt} represents the outcome of group j at time t; d_{jt} is the policy variable, so α is the main parameter of interest; θ_j is a time-invariant fixed effect for group j, while γ_t is a time fixed-effect; η_{jt} is a group x time error term that might be correlated over time, but uncorrelated across groups. Depending on the application, "groups" might stand for states, counties, countries, and so on. We assume that d_{jt} is nonstochastic.

There are N_1 treated groups and N_0 control groups. Let us assume that d_{jt} changes to 1 for all treated groups starting after date t^*. In this case, the DID estimator will be given by:

$$\begin{aligned}
\hat{\alpha} &= \frac{1}{N_1}\sum_{j=1}^{N_1}\left[\frac{1}{T-t^*}\sum_{t=t^*+1}^{T} Y_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*} Y_{jt}\right] - \frac{1}{N_0}\sum_{j=N_1+1}^{N}\left[\frac{1}{T-t^*}\sum_{t=t^*+1}^{T} Y_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*} Y_{jt}\right] \\
&= \alpha + \frac{1}{N_1}\sum_{j=1}^{N_1}\left[\frac{1}{T-t^*}\sum_{t=t^*+1}^{T} \eta_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*} \eta_{jt}\right] - \frac{1}{N_0}\sum_{j=N_1+1}^{N}\left[\frac{1}{T-t^*}\sum_{t=t^*+1}^{T} \eta_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*} \eta_{jt}\right] \\
&= \alpha + \frac{1}{N_1}\sum_{j=1}^{N_1} W_j - \frac{1}{N_0}\sum_{j=N_1+1}^{N} W_j
\end{aligned} \tag{2}$$

where $W_j = \frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\eta_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*}\eta_{jt}$.

It is clear from equation 2 that consistency of $\hat{\alpha}$ will depend on both $N_1 \to \infty$ and $N_0 \to \infty$. As shown in CT, if the number of treated groups (N_1) and the number of periods (T) are fixed, then the DID estimator

[7] The group x time DID aggregate model takes any individual level within group x time cell correlation in the errors into account (DL and Moulton (1986)). However, there might still be correlation of individuals in the same group at different periods in the aggregate model, as suggested by Bertrand et al. (2004).

is unbiased. However, this estimator is not consistent, since the first term, $\frac{1}{N_1}\sum_{j=1}^{N_1} W_j$, would not converge to zero when $N_0 \to \infty$.

The variance of the DID estimator, under the assumption that η_{jt} are independent across j, is given by:

$$var(\hat{\alpha}) = \left[\frac{1}{N_1}\right]^2 \sum_{j=1}^{N_1} var(W_j) + \left[\frac{1}{N_0}\right]^2 \sum_{j=N_1+1}^{N} var(W_j) \quad (3)$$

Note that the variance of the DID estimator is the sum of two components: the variance of the treated groups pre/post comparison and the variance of the control groups pre/post comparison. We allow for any kind of correlation between η_{jt} and $\eta_{jt'}$, which is captured in the linear combination of the errors W_j.

When there are many treated and control groups, Bertrand et al. (2004) suggest that CRVE at the group level works well, as this method allows for unrestricted intra-group and serial correlation in the residuals η_{jt}. One important point is that this method is not only cluster-robust, but also heteroskedasticity-robust. The CRVE has a very intuitive formula in the DID framework:[8]

$$\widehat{var(\hat{\alpha})}_{\text{Cluster}} = \left[\frac{1}{N_1}\right]^2 \sum_{j=1}^{N_1} \widehat{W}_j^2 + \left[\frac{1}{N_0}\right]^2 \sum_{j=N_1+1}^{N} \widehat{W}_j^2 \quad (4)$$

where $\widehat{W}_j = \frac{1}{T-t^*}\sum_{t=t^*+1}^{T} \hat{\eta}_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*} \hat{\eta}_{jt}$.

With CRVE, we calculate each component of the variance of the DID estimator separately. In other words, we use the residuals of the treated groups to calculate the component related to the treated groups, and the residuals of the control groups to calculate the component related to the control groups. This way, CRVE allows for unrestricted heteroskedasticity. While CRVE is very appealing when there are many treated and many control groups, equation 4 makes it clear why it becomes unappealing when there are few treated groups. In the extreme case when $N_1 = 1$, we will have $\widehat{W}_1^2 = 0$ *by construction*. Therefore, the variance of the DID estimator would be severely underestimated (MacKinnon and Webb (2015b)). The same problem applies to other clustered standard errors corrections such as BRL (Bell and McCaffrey (2002)). It is also problematic to implement heteroskedasticity-robust cluster bootstrap methods such as pairs-bootstrap and wild cluster bootstrap when there are few treated groups. In pairs-bootstrap, there is a high probability that the bootstrap sample will not include a treated unit. Wild cluster bootstrap generates variation in the residuals of each j by randomizing whether its residual will be $\hat{\eta}_{jt}$ or $-\hat{\eta}_{jt}$. However, in the extreme case with only one treated, this leads to $\widehat{W}_1 = 0$. Therefore, the wild cluster bootstrap would not generate

[8] Up to a degrees-of-freedom correction.

variation in the treated group. Another alternative presented by Bertrand et al. (2004) is to collapse the pre- and post-information. This approach would take care of the auto-correlation problem. However, in order to allow for heteroskedasticity, one would have to use heteroskedasticity-robust standard errors. In this case, this method would also fail when there are few treated groups.

It is clear, then, that the inference problem in DID models with few treated groups revolves around how to estimate the component of the DID estimator variance related to the treated group using the residuals $\hat{\eta}_{jt}$. Alternative methods use information on the residuals of the control groups in order to estimate the component of the variance related to the treated groups. These methods, however, rely on specific assumptions regarding the error terms. DL assume that the group x time errors are normal, homoskedastic, and serially uncorrelated. Under these assumptions, the variance of $\hat{\alpha}$ becomes:

$$var(\hat{\alpha}) \quad = \quad \frac{1}{NT}\frac{\sigma_\eta^2}{p(1-p)} \tag{5}$$

where $var(\eta_{jt}) = \sigma_\eta^2$ and p is the proportion of treated groups. Therefore, under these assumptions, one could recover the variance of $\hat{\alpha}$ by estimating σ_η^2 using the $T \times N$ estimated residuals $\hat{\eta}_{jt}$. As suggested by DL, if $T \times N$ is small, then one should compare the test statistic $t = \hat{\alpha}/\sqrt{\widehat{var(\hat{\alpha})}}$ to the student-t distribution instead of calculating the critical values based on the normal distribution. The assumption that errors are serially uncorrelated, however, might be unappealing in DID applications (Bertrand et al. (2004)).

CT provide an interesting alternative inference method that allows for unrestricted auto-correlation in the error terms. Their method uses the residuals of the control groups to estimate the distribution of the DID estimator under the null. The key difference relative to DL is that CT look at a linear combination of the residuals that takes into account any form of serial correlation instead of using the group x time level residuals. In the simpler case with only one treated group, $\hat{\alpha}-\alpha$ would converge to W_1 when $N_0 \to \infty$. In this case, they use $\{\widehat{W}_j\}_{j=2}^{N_0+1}$ (a linear combination of the control group residuals) to construct the distribution of W_1. While CT relax the assumption of no auto-correlation, it requires that errors are i.i.d. across groups, so that $\{\widehat{W}_j\}_{j=2}^{N_0+1}$ approximates the distribution of W_1 when $N_0 \to \infty$.

Finally, cluster residual bootstrap methods resample the residuals while holding the regressors constant throughout the pseudo-samples. The residuals are resampled at the group level, so that the correlation structure is preserved. It is possible that a treated group receives the residuals of a control group. While this helps when there are only few treated groups, a crucial assumption is that errors are homoskedastic. It is important to note that bootstrap alternatives with asymptotic refinements that focus on pivotal test statistics

would not work well in situations of few treated groups and heteroskedasticity. This happens because these methods require a consistent estimator of the variance. However, with N_1 fixed, the heteroskedasticity-robust methods to estimate the variance would not work properly.

2.2 The Heteroskedasticity Problem

As seen in Section 2.1, CRVE in DID models with few treated groups severely underestimates the variance of $\hat{\alpha}$. Alternative methods such as DL, CT and cluster residual bootstrap require strong distributional assumptions on the errors. In particular, they all require some kind of homoskedasticity. In this section, we show that these methods might not perform well in the presence of heteroskedasticity. In particular, we show that group x time DID aggregate models will be inherently heteroskedastic when there is variation in the number of observations per group and derive the implications of this heteroskedasticity for these inference methods.

We start with an individual-level DID model:

$$Y_{ijt} = \alpha d_{jt} + \theta_j + \gamma_t + \nu_{jt} + \epsilon_{ijt} \tag{6}$$

where Y_{ijt} represents the outcome of individual i in group j at time t; ν_{jt} is a group x time error term (possibly correlated over time), and ϵ_{ijt} is an individual-level error term. The main feature that defines a "group" in this setting is the assumption that errors $(\nu_{jt} + \epsilon_{ijt})$ of two individuals in the same group might be correlated, while errors of individuals in different groups are uncorrelated. For ease of exposition, we assume that ϵ_{ijt} are all uncorrelated, while allowing for unrestricted auto-correlation in ν_{jt}. However, our correction will require weaker assumptions on the error structure, as will be presented in Section 2.3.

When we aggregate by group x time, our model becomes the same as the one in equation 1:

$$Y_{jt} = \alpha d_{jt} + \theta_j + \gamma_t + \eta_{jt} \tag{7}$$

The important point is that errors in the group x time aggregate model (η_{jt}) are heteroskedastic across j, unless $M(j,t)$ is constant across j. More specifically:

$$\eta_{jt} = \nu_{jt} + \frac{1}{M(j,t)} \sum_{i=1}^{M(j,t)} \epsilon_{ijt} \tag{8}$$

where $M(j,t)$ is the number of observations in group t at time t. Therefore, assuming for simplicity that $M(j,t) = M_j$ is constant across j and T is fixed:

$$\begin{aligned} var(W_j) &= var\left(\frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\eta_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*}\eta_{jt}\right) \\ &= var\left(\frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\nu_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*}\nu_{jt} + \frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\left[\frac{1}{M_j}\sum_{i=1}^{M_j}\epsilon_{ijt}\right] - \frac{1}{t^*}\sum_{t=1}^{t^*}\left[\frac{1}{M_j}\sum_{i=1}^{M_j}\epsilon_{ijt}\right]\right) = \\ &= A + \frac{B}{M_j} \end{aligned} \tag{9}$$

for constants A and B, regardless of the auto-correlation of ν_{jt}.[9]

We are assuming so far that we have a panel of repeated cross-sections, so that ϵ_{ijt} are not correlated over time. If we had a panel and allow for the individual-level residuals to be auto-correlated, then we would have another term that would depend on the ϵ_{ijt} auto-correlation parameter divided by the number of observations, so we would still end up with the same formula, $var(W_j) = A + \frac{B}{M_j}$.

This heteroskedasticity in the error terms of the aggregate model implies that, when the number of observations in the treated groups are (large) small relative to the number of observations in the control groups, we would (over-) underestimate the component of the variance related to the treated group when we estimate it using information from the control groups. This implies that inference methods that do not take that into account would tend to (under-) over-reject the null hypothesis when the number of observations of the treated groups is (large) small.

Note that, if $A > 0$, this would not be a problem when $M(j,t) \to \infty$. In this case, $var(W_j) \to A$ for all j. In other words, when the number of observations in each group x cell is large, then the correlated part of the error would dominate. In this case, if we assume that the group x time error ν_{jt} is i.i.d., then $\frac{var(W_j)}{var(W_j')} \to 1$, which implies that the residuals of the control groups would be a good approximation for the distribution of the treated groups error even when the number of observations in each group is different. This is one of the main rationales used in DL to justify the homoskedasticity assumption in the aggregate model.

However, an interesting case occurs when $A = 0$. In this case, even though $var(W_j) \to 0$ for all j when $M_j \to \infty$, the ratios $\frac{var(W_j)}{var(W_{j'})}$ remain constant (unless $\frac{M_j}{M_{j'}} \to 1$), which implies that the aggregate model remains heteroskedastic even asymptotically. Therefore, CT, DL and cluster residual bootstrap would tend to (under-) over-reject the null hypothesis when the number of observations of the treated groups are (large)

[9] When the number of observations per group is not constant over time, the formula will be: $var(W_j) = \tilde{A} + \tilde{B}\left[\left(\frac{1}{T-t^*}\right)^2 \sum_{t=t^*+1}^{T}\frac{1}{M(j,t)} + \left(\frac{1}{t^*}\right)^2 \sum_{t=1}^{t^*}\frac{1}{M(j,t)}\right]$, for constants $\tilde{A}$ and $\tilde{B}$.

small relative to the number of observations of the control groups even when there is a large number of individual observations.

2.3 Inference with Heteroskedasticity Correction

As discussed in Section 2.1, the main challenge in estimating the variance of $\hat{\alpha}$ when there are few treated groups is how to estimate the component related to the treated groups. The CRVE estimates this component of the variance without using information from the control groups. While this approach has the appealing property of allowing for unrestricted heteroskedasticity, it is unfeasible when the number of treated groups is small. On the other extreme, other methods surpass the problem of few treated groups by using information from the control groups. The problem with these approaches is that they require homoskeadsticity.

In this section, we derive an inference method that uses information from the control groups to estimate the variance of the treated groups while allowing for heteroskedasticity. Our approach assumes that we know how the heteroskedasticity is generated, which is the case when heteroskedasticity is generated by variation in the number of observations per group. Under this assumption, we can re-scale the residuals of the control groups using the (estimated) structure of the heteroskedasticity in a way that allows us to use this information to estimate the distribution of the error for the treated groups. Our method only requires information on the heteroskedasticity structure for a linear combination of the errors, which implies that we do not have to impose strong assumptions on the structure of the serial correlation of the errors. While we motivate our methods based on heteroskedasticity generated by variation in the number of groups, it is important to note that our method is more general.

Our method is an extension of the cluster residual bootstrap with H_0 imposed where we correct the residuals for heteroskedasticity. In cluster residual bootstrap with H_0 imposed, we estimate the DID regression imposing that $\alpha = 0$, generating the residuals $\{\widehat{W}_j^R\}_{i=1}^N$. If the errors are homoskedastic, then, under the null, $\widehat{W}_j^R$ would have the same distribution across j, which implies that we can resample with replacement $\mathcal{B}$ times from $\{\widehat{W}_j^R\}_{i=1}^N$, generating $\{\widehat{W}_{j,b}^R\}_{i=1}^N$. Then we can calculate our bootstrap estimates as $\hat{\alpha}_b = \frac{1}{N_1}\sum_{j=1}^{N_1}\widehat{W}_{j,b}^R - \frac{1}{N_0}\sum_{j=N_1+1}^{N}\widehat{W}_{j,b}^R$. Note that, in our setting, we do not need to work with the group x time residuals $\hat{\eta}_{jt}$ to construct our bootstrap estimates. Instead, we can work with a linear combination of the residuals that takes into account any form of auto-correlation in the residuals. This is one of the key insights of CT.

As explained in Section 2.1, the problem with cluster residual bootstrap is that it requires the residuals to be homoskedastic. In Theorem 1 in Appendix A, we show that, if we know the variance of each random vari-

able W_j, then we can re-scale the residuals $\widehat{W}^R_{j,b}$ and use a cluster residual bootstrap on the re-scaled residuals even if the model is heteroskedastic. First, we normalize each observed $\widehat{W}_{j'}$ by $\widehat{W}^{norm}_{j'} = \widehat{W}^R_{j'} \frac{1}{\sqrt{var(W_{j'})}}$. Then we generate a bootstrap sample with the re-scaled residuals $\widetilde{W}_{j,b} = \widehat{W}^{norm}_{j,b} \sqrt{var(W_j)}$. As a result, this procedure generates bootstrap estimators $\hat{\alpha}_b = \frac{1}{N_1} \sum_{j=1}^{N_1} \widetilde{W}_{j,b} - \frac{1}{N_0} \sum_{j=N_1+1}^{N} \widetilde{W}_{j,b}$ with the same distribution as the DID estimator. The main assumption we need is that $\{W_j\}_{j=1}^N$, which is a linear combination of the error terms η_{jt}, are independent across j and have the same distribution up to the variance parameter.[10] As in CT, it is important to note that we only need the variance of a linear combination of the errors. This point is crucial for our method, because we do not need to know the serial correlation structure of the errors η_{jt}.

The main problem, however, is that $var(W_j)$ is generally unknown, so it needs to be estimated. In Theorem 2 in Appendix A, we show that this heteroskedasticity correction works asymptotically when $N_0 \to \infty$ if we have a consistent estimator for $var(W_j)$. Our method assumes that we know the structure of the heteroskedasticity. In our setting, we assume that $var(W_j)$ is a function that depends only on M_j, $G(M_j) = A + \frac{B}{M_j}$, for constants A and B. The error structure we assumed in Section 2.2 implies this structure. However, this assumption is more general. In particular, we do not have to make any assumption on the auto-correlation of η_{jt}. Given this assumption, we can run a regression of $\hat{W}_j^2$ on $\frac{1}{M_j}$ and a constant, and then use the predicted $\widehat{G(M_j)}$ to generate $\widetilde{\widehat{W}}_{j,b} = \widehat{W}^R_{j,b} \sqrt{\frac{\widehat{G(M_j)}}{\widehat{G(M_{j,b})}}}$.[11] We show in Theorem 3 in Appendix A that $\widehat{G(M_j)}$ is a consistent estimator for $var(W_j)$. Note that we do not need individual-level data to apply this method, provided that we have information on the number of observations that were used to calculate group x time averages.

Finally, a problem with cluster bootstrap methods when there are few clusters is that there will be few possible combinations of bootstrap samples (Cameron et al. (2008), Webb (2014), and MacKinnon and Webb (2015a)). To ameliorate this problem, we apply the idea of wild cluster bootstrap to our method. Therefore, for each j, we sample either $\widetilde{\widehat{W}}_{j,b}$ with probability 0.5 or $-\widetilde{\widehat{W}}_{j,b}$ with probability 0.5. This procedure provides a smoother bootstrap distribution.

Summarizing, our bootstrap procedure consists of:

[10] Note that this assumption is weaker than assuming that the sequences $\{\eta_{j1}, ..., \eta_{jT}\}$ are independent and have the same distribution up to a variance parameter across j.

[11] When the number of observations per group is not constant over time, we regress $\widehat{W}_j^2$ on $\left[\left(\frac{1}{T-t^*}\right)^2 \sum_{t=t^*+1}^{T} \frac{1}{M(j,t)} + \left(\frac{1}{t^*}\right)^2 \sum_{t=1}^{t^*} \frac{1}{M(j,t)}\right]$ and a constant.

1. Calculate the DID estimate:

$$\hat{\alpha} = \frac{1}{N_1}\sum_{j=1}^{N_1}\left[\frac{1}{T-t^*}\sum_{t=t^*+1}^{T} Y_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*} Y_{jt}\right] - \frac{1}{N_0}\sum_{j=N_1+1}^{N}\left[\frac{1}{T-t^*}\sum_{t=t^*+1}^{T} Y_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*} Y_{jt}\right]$$

2. Estimate the DID model with H_0 imposed ($Y_{jt} = \alpha_0 d_{jt} + \theta_j + \gamma_t + \eta_{jt}$), and obtain $\{\widehat{W}_j^R\}_{i=1}^N$. Usually the null will be $\alpha_0 = 0$.

3. Estimate $G(M)$ by regressing $\left(\widehat{W}_j^R\right)^2$ on a constant and $\frac{1}{M_j}$.

4. Use $\widehat{G(M)}$ to obtain the normalized residuals $\widehat{W}_{j'}^{norm} = \widehat{W}_{j'}^R \frac{1}{\sqrt{\widehat{G(M_{j'})}}}$

5. Do $\mathcal{B}$ iterations of this step. On the b^{th} iteration:

   (a) Resample with replacement N times from $\{\widehat{W}_j^{norm}\}_{i=1}^N$ to obtain $\left\{\widetilde{\widehat{W}}_{j,b}\right\}_{i=1}^N$, where $\widetilde{\widehat{W}}_{j,b} = \widehat{W}_{j,b}^{norm}\sqrt{\widehat{G(M_j)}}$ with probability 0.5 and $-\widehat{W}_{j,b}^{norm}\sqrt{\widehat{G(M_j)}}$ with probability 0.5.

   (b) Calculate $\hat{\alpha}_b = \frac{1}{N_1}\sum_{j=1}^{N_1}\widetilde{\widehat{W}}_{j,b} - \frac{1}{N_0}\sum_{j=N_1+1}^{N}\widetilde{\widehat{W}}_{j,b}$.

6. Reject H_0 at level a if and only if $\hat{\alpha} < \hat{\alpha}_b[a/2]$ or $\hat{\alpha} > \hat{\alpha}_b[1-a/2]$, where $\hat{\alpha}_b[q]$ denotes the q^{th} quantile of $\hat{\alpha}_1, ..., \hat{\alpha}_{\mathcal{B}}$.

2.4 Randomization Inference and Permutation Tests

We assume in our model that treatment assignment is nonstochastic, while the stochastic elements in the model come from η_{jt}, ν_{jt}, and ϵ_{ijt}. This departs crucially from Randomization Inference (RI), which considers that the only stochastic component of the model is the treatment assignment (Fisher (1935)). In RI, one calculates the test statistic under all possible combinations of treatment assignment, and rejects the null if the observed realization in the actual experiment is extreme enough. This idea is closely related to CT. In fact, CT propose an alternative way to implement their method which is *heuristically* motivated by the literature on permutation tests and RI. As stated in Lehmann and Romano (2008), RI provides exact test statistics based solely on the null of no treatment effects and the fact that treatment was randomly assigned, not depending on any assumption regarding the characteristics of outcome, covariates and residuals. Young (2015) argues that many published papers in Economics that use standard inference methods in randomized experiments produce invalid testing, and proposes the use of RI methods. While we agree that RI provides a powerful inference method in randomized experiments, we believe the assumption that the only stochastic

element is the allocation of treatment is unreasonable in many DID applications. Still, it is worth contrasting the RI solution to our method.

Imagine first that one does not have information on the number of observations per group (or, more generally, on the variable that generates heteroskedasticity). If one assumes that treatment was randomly assigned, then all the hypotheses for RI would be satisfied, even if we have groups of different sizes. Small groups would have more variable outcomes, which implies that one would reject more often when the treatment is assigned to these groups (as explained in 2.2) but, unconditionally, one would still have a test with the correct size. Intuitively, the residuals η_{jt} would depend on the group sizes (it would attain more extreme values for smaller groups), but this is not a problem for RI because this method works regardless of the characteristics of the residuals. One interesting point is that one would see more statistically significant results when the treated group is small, which is actually the case when the estimator should be *less* precise. However, this does not invalidate the test, as it would continue to (unconditionally) reject with the correct size under the null.

If one does have information on group sizes, however, then an unconditional permutation test would not be correct. As argued by Yates (1984), a permutation test should incorporate all the available information. Once one knows that the treated groups are (large) small relative to the control groups, then one knows that a permutation test that ignores this information would (under-) over-reject the null when the null is true. Therefore, this test would no longer have the correct size. There are at least two ways of incorporating this information in a permutation test. One would be to apply the permutation conditional on the information on group sizes. However, if there are few control groups with the same size as the treated groups, then one would not have many possible permutations. In the particular case where there is one treated group and no control group of the same size, this conditional permutation test would generate a p-value interval of $[0, 1]$. Another alternative would be to use a test statistic that does not depend on the size of the groups, as suggested by Canay et al. (2014). For example, MacKinnon and Webb (2015a) suggest a permutation test on a t-statistic, which is constructed using CRVE. Their method works when the numbers of treated and control groups are large enough, as asymptotically the t-statistic would have the same distribution under the null for all permutations. However, their method does not work well with very few treated groups. In particular, their method collapses to CT when there is only one treated group. The reason is that CRVE would assign an estimated variance for the treated group equal to zero, so there would not be much variation in the *estimated* variance of the placebo estimators. The key point is that we go back to the original problem of estimating the variance of the treated groups using CRVE with few treated groups. In contrast, our

method provides a valid correction for heteroskedasticity even when there is only one treated group.

2.5 Alternative Estimation Methods - Synthetic Control

The Synthetic Control estimator was proposed by Abadie and Gardeazabal (2003) and Abadie et al. (2010) to deal with situations where there is only one treated group. This method extends the traditional DID framework by using a data-driven procedure to construct a suitable comparison group. The main idea is to use the pre-treatment period to construct a counterfactual for the treated group given by $\hat{Y}_{1t}^N = \sum_{j=2}^{N_0+1} \hat{\omega}_j Y_{jt}$, where the weights $\hat{\omega}_j$ are estimated so that the differences between actual and estimated pre-treatment outcomes (Y_{1t} and $\hat{Y}_{1t}^N$) and covariates (X_{1t} and $\hat{X}_{1t}^N$) are minimized.[12] In the Synthetic Control approach, one needs to decide which variables to include to estimate the weights $\hat{\omega}_j$. Particularly important for our application, one can either include the Y_{jt} for all pre-treatment t, or leave some of the pre-treatment Y_{jt} out.

The inference method suggested in Abadie et al. (2010) is a permutation test where one estimates placebo regressions using each of the control units as a placebo treatment. In essence, this is the same as what CT do in the DID framework. However, one important difference relative to permutation tests on the treatment parameter is that Abadie et al. (2010) suggest that one should look at the ratio of post-/pre-treatment Mean Squared Predicted Error (MSPE). One of their motivations to look at this ratio is to obviate the necessity of excluding placebo runs that did not provide a good fit prior to the treatment. For example, if the outcome variable of one placebo group is always lower than the outcome variables of the other groups, then the estimated counterfactual outcome for this group would always be atypically higher than the actual outcome, both before and after the treatment. Therefore, when one divides by the pre-treatment MSPE, this corrects for the fact that the Synthetic Control estimators for this placebo group will always be large. We show now that, under some circumstances, this inference method corrects for heteroskedasticity. We derive the conditions under which this is the case.

Consider the model in Abadie et al (2010),

$$Y_{jt} = \alpha_{1t} d_{it} + \gamma_t + \beta_t Z_j + \lambda_t \mu_j + \eta_{jt}^{SC} \tag{10}$$

where d_{jt} is an indicator variable that equals one if j is the treated region and $t > T_0$ (pre-intervention period), and Z_j is a vector of observed covariates for region j. The unobserved residual is $u_{jt} = \lambda_t \mu_j + \eta_{jt}^{SC}$. They assume that the η_{jt}^{SC} are $i.i.d$ cross j and t, and that η_{jt}^{SC} are mean indepedent of $\{Z_j, \mu_j\}_{j=1}^N$. We want to

[12] For more details, see Abadie et al. (2010).

show that in some cases, looking at this ratio provides proper hypothesis testing under heteroskedasticity. For simplicity, consider that we have three periods, two before the treatment and one after the treatment. Suppose that we construct our Synthetic Control estimator using only the outcome variable from period 1. Under the Synthetic Control assumptions the difference $Y_{11} - \hat{Y}_{11}^N$ will be close to zero, since the weights used to construct $\hat{Y}_{11}^N$ were chosen to minimize this difference. In addition, using Abadie et al (2010) derivations, for $t \in \{2, 3\}$ we have that:

$$\begin{aligned} Y_{1t} - \hat{Y}_{1t}^N &= \alpha_{1t} d_{it} + \textstyle\sum_{j=2}^{N_0+1} w_j^* \sum_{s=1}^{T_0} \lambda_t \left(\sum_{n=1}^{T_0} \lambda_n' \lambda_n \right)^{-1} \lambda_s' \left(\eta_{js}^{SC} - \eta_{1s}^{SC} \right) + \\ &\quad - \textstyle\sum_{j=2}^{N_0+1} w_j^* \left(\eta_{jt}^{SC} - \eta_{1t}^{SC} \right) \end{aligned} \tag{11}$$

Therefore, under the null that $\alpha_{1t} = 0$:

$$\begin{aligned} E[Y_{1t} - \hat{Y}_{1t}^N]^2 &= \sum_{j=2}^{N_0+1} w_j^{*2} \cdot \left(\sum_{s=1}^{T_0} \lambda_t \left(\sum_{n=1}^{T_0} \lambda_n' \lambda_n \right)^{-1} \lambda_s' \right)^2 Var[\eta_{js}^{SC} - \eta_{1s}^{SC}] + \\ &\quad + \sum_{j=2}^{N_0+1} w_j^{*2} Var\left[\eta_{jt}^{SC} - \eta_{1t}^{SC} \right], \text{ for } t \in \{2, 3\} \end{aligned} \tag{12}$$

The key point is that, under the assumption that $\{\eta_{jt}^{SC}\}_{t=1}^T$ is identically distributed across t, then equation 12 will not depend on t. Therefore, the post-/pre-intervention RMSE ratio, $\frac{E[Y_{13} - \hat{Y}_{13}^N]^2}{E[Y_{12} - \hat{Y}_{12}^N]^2}$, will be equal to one. This will also be true in the permutation when we consider group j as treated, even if $var(\eta_{1t}^{SC}) \neq var(\eta_{jt}^{SC})$. This is why the inference method proposed by Abadie et al. (2010) corrects the information from the control groups variation so that it becomes comparable to the variation in the treated group.[13] Note that this assumption on the error structure is stronger than the structure we need for our method. In particular, the inference method proposed by Abadie et al. (2010) requires that residuals are independent across time, while our method allows for unrestricted serial correlation in the residuals. Therefore, the synthetic control inference method fails to correct for heteroskedasticity if the sample in the pre-treatment is smaller or larger than the sample in the post-treatment, even if the ratio of number of observations across groups remains constant.

Another case in which the synthetic control inference approach would not to correct for heteroskedasticity is when there is only one pre-treatment period. In this case, one would have to estimate the weights using the single pre-treatment period. One could still calculate the RMSE ratio, since $Y_{j1} - \hat{Y}_{j1}^N$ will not be identical

[13] This argument would remain valid if we had more than one post period and/or more than one pre period not included in the estimation of ω_j.

to zero. However, this division would not re-scale the numerator correctly. The same problem applies when we have more than one pre-treatment period but include all pre-treatment periods to estimate the weights. Finally, it is also important to note that the permutation graphical analyses in Abadie et al. (2010) would still suffer from the heteroskedasticity problem we highlight in this paper.[14] An easy way to fix this problem is to divide each placebo estimate by the squared root of its pre-treatment RMSE and multiply it by the squared root of the the pre-treatment RMSE of the treated group.

3 Monte Carlo Evidence

In this section we provide Monte Carlo evidence of different hypothesis testing methods in DID. We also simulate the inference method for Synthetic Control models proposed by Abadie et al. (2010) in Section 3.2. We assume that the underlying data generating process (DGP) is given by:

$$Y_{ijt} = \nu_{jt} + \epsilon_{ijt} \tag{13}$$

In most simulations, we estimate a DID model given by equation 6 where only $j = 1$ is treated and $T = 2$, and then we test the null hypothesis of $\alpha = 0$ using different hypothesis testing methods. We consider variations in the DGP along three dimensions:

1. The number of groups: $N_0 + 1 \in \{25, 50, 100, 400\}$.
2. The intra-group correlation: ν_{jt} and ϵ_{ijt} are drawn from normal random variables. We hold constant the total variance $var(\nu_{jt} + \epsilon_{ijt}) = 1$, while changing $\rho = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\epsilon^2} \in \{.01\%, 1\%, 4\%\}$.
3. The number of observations within group: we draw for each group j the number of observations per period from a discrete uniform random variable with range $[\underline{M}, \overline{M}] \in \{[50, 200], [200, 800], [50, 950]\}$.[15]

For each case, we simulated 100,000 estimates. We present rejection rate results for inference using robust standard errors in the individual-level OLS regression, CT, DL, and for the cluster residual bootstrap with and without our heteroskedasticity correction. We do not include in the simulations methods that allow for unrestricted heteroskedasticity. As explained in Section 2.1, these methods do not work well when there is only one treated group. We also do not include MacKinnon and Webb (2015a) method in the simulations because their method collapses to CT when there is only one treated group.

[14] Figures 4 to 7 in Abadie et al. (2010).

[15] In the Monte Carlo simulations, we always consider the case $M(j,t) = M_j$. In the simulations with real datasets in Section 4, there is variation in $M(j,t)$ across t.

3.1 Inference in DID Models

3.1.1 Test Size

We present in Table 1 results from simulations using 400 groups (one treated and 399 controls) for different numbers of observations per group and for different values of the intra-group correlations. In panel A, we present results when the number of individual observations per group varies from 50 to 200. Column 1 shows that average rejection rates for a test with 5% significance using robust standard errors in the individual level DID regression. The rejection rate is slightly higher than 5% when the intra-group correlation $\rho = 0.01\%$ (5.4%), but increases sharply for larger values of the intra-group correlation. Rejection rate is 19% when $\rho = 1\%$ and 42% when $\rho = 4\%$. When we use DL, CT or cluster residual bootstrap without correction, average rejection rate is always around 5% (columns 3, 5, and 7). However, this average rejection rate hides an important variation with respect to the number of observations in the treated group (M_1).

In Figure 1.A, we show rejection rates for cluster residual bootstrap without correction conditional on the size of the treated group.[16] The rejection rate is around 14% when the treated group is in the first decile of number of observations per group, while it is only 0.8% when the treated group is in the 10th decile. Note also that this distortion in rejection rates is not confined to the extremes of the distribution of group sizes. For example, the rejection rate is 3% when the treated group is in the 6th decile of number of observations per group. We summarize this variation in rejection rates by looking at the absolute difference in rejection rates for each decile of M_1 relative to the average rejection rate. Then we average these absolute differences across deciles. We present these results in columns 4, 6, and 8 for the methods without heteroskedasticity correction. Conditional on the number of observations of the treated group, these methods present an average variation in the rejection rates of 3.4-3.9 percentage points for a 5% significance test.

We present rejection rates by decile of the treated group for cluster residual bootstrap without correction when $\rho = 1\%$ and when $\rho = 4\%$ in Figures 1.B and 1.C, respectively. As expected, this variation in rejection rates becomes less relevant when the intra-group correlation becomes stronger. This happens because the aggregation from individual to group x time averages induces less heteroskedasticity in the residuals when a larger share of the residual is correlated within group. Still, even when $\rho = 4\%$ the difference in rejection rates by number of observations in the treated group remains relevant. The rejection rate is around 6.5% when the treated group is in the first decile of number of observations per group, while it is 4.2% when the treated group is in the 10th decile. The average absolute difference in rejection rates for DL, CT and for the

[16] Results for DL and CT are similar.

residual bootstrap without correction is around 0.7 percentage points in this scenario.

Given that inference using these methods is problematic when there is variation in the number of observations per group, we consider our residual bootstrap method with heteroskedasticity correction derived in Section 2.3. We present rejection rates by decile of the treated group when the intra-group correlation is 0.01%, 1% and 4% in Figures 1.D to 1.F. Average rejection rates using our method are always around 5% and, more importantly, there is no variation with respect to the number of observations in the treated group. These results are also presented in columns 9 and 10 of Table 1. The average absolute difference in rejection rates is only around 0.1-0.2 percentage points, regardless of the value of the intra-group correlation.

In panel B of Table 1 we present the simulation results when the number of observations per group increases from $[50, 200]$ to $[200, 800]$. We increase the number of observations per group while holding the ratio between the number of observations in different groups constant. Note that increasing the number of observations per group worsens the over-rejection problem of inference relying in robust OLS standard errors. When we consider DL, CT and residual bootstrap without correction, increasing the number of observations per group ameliorates the problem of (over-) under-rejecting the null when M_1 is (small) large relative to the number of observations in the control groups. In particular, when $\rho = 4\%$ the average absolute difference in rejection rates across deciles of M_1 is only 0.3 percentage points. However, increasing the number of observations has no detectable effect when the intra-group correlation is 0.01%. This happens because in this case the individual component of the residual becomes more relevant. Therefore, the ratio between the variance of W_1 and the variance of W_j becomes less sensitive with respect to the number of observations per group. As explained in Section 2, in the extreme case with $\rho = 0$, heteroskedasticity would still be a problem even when $M \to \infty$.

In panel C of Table 1, we present the simulation results when the number of observations vary from 50 to 950. Therefore, the average number of observations remains constant, but we have more variation in M relative to the simulations in panel B. As expected, more variation in the number of observations per group worsens the inference problem we highlight in CT, DL and residual bootstrap without correction. On the contrary, our residual bootstrap with heteroskedasticity correection remains accurate irrespective of the variation in the number of observations per group.

As presented in Section 2.3, our method works asymptotically when $N_0 \to \infty$. This assumption is important for two reasons. First, as in any other cluster bootstrap method, a small number of groups implies a small number of possible distinct pseudo-samples. In this case, the bootstrap distribution will not be smooth even with many bootstrap replications (Cameron et al. (2008)). In order to mitigate this problem,

we apply the insight of wild cluster bootstrap to our method, so that we can generate more variation in the bootstrap samples, as explained in Section 2.3. Additionally, our method requires that we estimate $G(M)$ using the group x time aggregate data so that we can apply our heteroskedasticity correction. If there are only a few groups, then our estimator of $G(M)$ will be less precise. In particular, it might be the case that $\widehat{G(M_j)} < 0$ for some j, which implies that we would not be able to normalize the residual of observation j. When $\widehat{G(M_j)} < 0$ for some j, we used the following rule: if $\hat{A} < 0$, then we used $\widehat{G(M_j)} = \frac{1}{M_j}$, as $\hat{A} < 0$ would suggest that there is not not a large intra-group or serial correlation problem. If $\hat{B} < 0$, then we used $\widehat{G(M_j)} = 1$, as $\hat{B} < 0$ would suggest that there is not much heteroskedasticity. It is important to note that asymptotically this rule would not be relevant, since $G(M) > 0$ for all M. We had $\widehat{G(M_j)} > 0$ for all j in more than 99.97% of our simulations with $N = 400$. However, when there are fewer control groups, the function $G(M)$ will be estimated with less precision.

We present in Tables 2 to 4 and in Figures 2 to 4 the simulation results when the total number of groups are 100, 50 and 25. Average rejection rates are always lower than 5.3% when the total number of groups is 100 or 50, which is reasonably close to the correct size of the test. More importantly, the average absolute difference in rejection rates is always lower than 0.5 percentage points, suggesting that there is not much variation in rejection rates depending on the size of the treated group. These results are confirmed in Figures 2 and 3. When we have 25 groups, then average rejection rates are slightly higher, at around 5.5%, and we start to have more variation depending on the size of the treated group. As shown in Figure 4, there is some distortion in rejection rates when the treated group is in the first decile of group size. Still, our method provides reasonably accurate hypothesis testing with 25 groups. In particular, our method provides substantial improvement relative to alternative methods when the intra-group correlation is not too strong.

3.1.2 Test Power

We have focused so far on Type I error. We saw in Section 3.1.1 that our method is efficient in providing tests that reject the null with the correct size when the null is true. We are interested now in whether our tests have power to detect effects when the null hypothesis is false. We run the same simulations as in Section 3.1.1, with the difference that we now add an effect of β standard deviations for observation $\{ijt\}$ when $d_{jt} = 1$. Given that we know the DGP in our Monte Carlo simulations, we can calculate the variance of $\hat{\alpha}$ given the parameters of the model, so that we can generate a t-statistic $t = \frac{\hat{\alpha}}{\sigma_{\hat{\alpha}}}$. Using Neyman-Pearson Lemma, since the errors in our DGP are normally distributed, we know that a test based on this t-statistic is the uniformly most powerful test (UMP). We then compare the power of the bootstrap with

our heteroskedasticity correction with the power of the UMP test.

In Figure 5, we present power results for different intra-group correlation parameters and for different distributions of group sizes when there are 400 groups (1 treated and 399 control groups) separately when the treated group is above and below the median of number of observations per group. The most important feature in these graphs is that the power of our method converges to the power of the UMP test when we have many control groups in all intra-group correlation and group size scenarios. It is also interesting to note that the power is higher when the treated group is larger. This is reasonable, since the main component of the variance of the DID estimator with few treated and many control groups comes from the variance of the treated groups. The difference in power for above- and below-median treated groups vanishes when the intra-group correlation increases. This happens because a higher intra-group correlation makes the model less heteroskedastic. Finally, the power of the test decreases with the intra-group correlation which reflects that, for a given number of observations per group, a higher intra-group correlation implies more volatility in the group x time regression.

When we have 25 groups (1 treated and 24 control), then the power of our method is slightly lower than the power of the UMP test (Figure 6). This is partially explained by fact that we need to estimate the function $G(M)$ and, with a finite number of control groups, this function would not be precisely estimated. Still, the power of our method is relatively close to the power of the UMP test, especially when the intra-group correlation is not high.

3.2 Inference in Synthetic Controls

An alternative estimation method when there is only one treated group is to use the Synthetic Control Estimator. As explained in Section 2.5, one inference method suggested in Abadie et al. (2010) calculates the ratio of post-/pre-treatment RMSE of the Synthetic Control Estimator and compares it to the same ratio when we use the control groups as placebo treatments. We present in Figures 7.A to 7.C rejection rates for the case with $T = 2$, with one pre- and one post-intervention periods. We consider the case with $N = 50$ and $M \in [50, 950]$. The average rejection rates is 6%, which simply reflects that p-values in permutation tests with few groups are not point identified. We are more interested in how rejection rates vary with the size of the treated group. When $\rho = 0.01\%$, rejection rates are higher when the treated group is small (Figure 7.A). This happens because the post-treatment RMSE used in the numerator is higher when the treated group is smaller, due to the heteroskedasticity generated by the variation in the number of observations per group. However, the pre-treatment RMSE used in the denominator is just an error term reflecting the fact

that Y_{11}^N will not be identical to Y_{11}, so the ratio will decrease with M_1. When ρ is higher, a given variation in the number of observations per group generates less heteroskedasticity, so this effect is weaker (Figures 7.B and 7.C). Exactly the same pattern happens in Panel B, where we simulate a case with $T = 3$ with 2 pre-treatment periods, but include both Y_{j1} and Y_{j2} to estimate the weights (Figures 7.D to 7.F).

In Figures 7.G to 7.I, we consider again the case with $T = 3$ periods, but now we use only the first period to estimate the weights. In this case, the pre-treatment RMSE used in the denominator is higher when the treated group is smaller, since it includes the predicted error related to the pre-treatment period $t = 2$. As explained in Section 2.5, while both the numerator and the denominator decrease with M, the ratio will be constant under the assumption that the residuals are i.i.d. across time and i.i.d across groups up to the variance parameter (note that our heteroskeadsticity correction method allows for unrestricted auto-correlation across time within group). This implies that the difference in rejection rates for small and large groups is corrected using this inference method. The only qualification is that rejection rates are slightly *lower* when the treated group is small. This happens because when the treated group is small, it is more likely that it will not be possible to provide a good fit for the treated group. In this case, the pre-treatment RMSE will be larger. Again, this problem is less relevant when ρ is larger, since this implies that variation in M generates less heteroskedasticity.

4 Simulations with Real Datasets

The results presented in Section 3 suggest that heteroskedasticity generated by variation in group sizes invalidates inference methods that rely on homoskedasticity such as DL, CT and cluster residual bootstrap, while our method performs well in correcting for heteroskedasticity when there are 25 or more groups. However, a natural question that arises is whether these results are "externally valid." In particular, we want to know (i) whether heteroskedasticity generated by variation in group sizes is a problem in real datasets with large number of observations, and (ii) whether our method works in real datasets, where we do not have control over the DGP. More specifically, our DGP implies that the *real* variance of W_j would have exactly the relationship $var(W_j) = A + \frac{B}{M_j}$, which might not be the case in real datasets. To illustrate the magnitude of the heteroskedasticity problem and to test the accuracy of our method, we conduct simulations of placebo interventions using two different real datasets: the American Community Survey (ACS) and the Current Population Survey (CPS).

We consider two different group levels for the ACS based on the geographical location of residence: Public

Use Microdata Areas (PUMA) and states. Simulations using placebo interventions at the PUMA level would be a good approximation to our assumption that N_1 is small while $N_0 \to \infty$. Simulations using placebo interventions at the state level would mimic situations of DID designs that are commonly used in applied work where the treatment unit is a state, with a dataset that includes a very large number of observations per group x time cell. We also consider the CPS for simulations with more than two periods. As shown in Bertrand et al. (2004), this dataset exhibits an important serial correlation in the residuals, so we want to check whether our method method is efficient in correcting for that.

We use the ACS dataset for the years 2005 to 2013, and the CPS Merged Outgoing Rotation Groups for the years 1979 to 2014. We extract information on employment status and earnings for women between ages 25 and 50, following Bertrand et al. (2004). We present in Table 5 the distribution of number of observations per group x cell for the PUMA-level ACS (column 1), for the state-level ACS (column 2) and for the state-level CPS (column 3). There are, on average, 778 observations in each PUMA x time cell in the ACS. This number, however, hides an important heterogeneity in cell sizes. The 10th percentile of PUMA x time cell sizes is 174, while the 90th percentile is 1,418. There is also substantial heterogeneity in state x time cell sizes in the ACS. While the average cell size is 10,138, the 10th percentile is 1,250, while the 90th percentile is 21,099. Finally, the state x time cells in the CPS have substantially fewer observations compared to the ACS. While the average cell size is 771, the 10th percentile is 392, while the 90th percentile is 1709.

For the ACS simulations, we consider pairs of two consecutive years and estimate placebo DID regressions using one of the groups (PUMA or state) at a time as the treated group. Note that this differs from Bertrand et al. (2004) simulations, as they randomly selected half of the states to be treated. In each simulation, we test the null hypothesis that the "intervention" has no effect ($\alpha = 0$) using robust standard errors, and bootstrap with and without our heteroskedasticity correction. Since we are looking at placebo interventions, if the inference method is correct, then we would expect to reject the null roughly 5% of the time for a test with 5% significance level. For each pair of years, the number of PUMAs that appear in both years ranges from 427 to 982, leading to 5,188 regressions in total. For the state-level simulations, we have $51 \times 8 = 408$ regressions (we include Washington, D.C.). For the CPS simulations, we used 2, 4, 6 or 8 consecutive years always using the first half of the years as pre-treatment and the other half as post-treatment. This leads to 1479 to 1785 regressions, depending on the number of years used in each regression.

4.1 American Community Survey (ACS) Results

In Panel A of Table 6, we present results from simulations using the PUMA-level treatments using the ACS. In column 1, we show rejection rates using OLS robust standard errors in the individual-level DID regression. Rejection rates for a 5% significance test are 7.2% when the outcome variable is employment, and 8.1% when it is log wages. This over-rejection suggests that there is important intra-group correlation that the robust individual-level standard error does not take into account. In column 3 of Table 6, we present results for the bootstrap without the heteroskedasticity correction (results for DL and CT are simular). As in the Monte Carlo simulations, average rejection rates without correction are very close to 5%. However, there is substantial variation when we look at rejection rates conditional on the size of the treated group. We present in column 4 of Table 6 the difference in rejection rates when the number of observations in the treated group is above and below the median.[17] For both outcome variables, the rejection rate is 8 percentage points lower when the treated group has a group size above the median. This implies a rejection rate of almost 9% when the treated group is below the median, and slightly lower than 1% when the treated group is above the median. In columns 5 and 6 of Table 6, we present the rejection rates using bootstrap with our heteroskedasticity correction. For both outcomes, average rejection rate has the correct size of 5% and, more importantly, there is virtually no difference between rejection rates when the treated group is above or below the median. Therefore, our method was successful in correcting for the heteroskedasticity problem even in a setting where we do not have control over the DGP.

We present in Panel B of Table 6 the results for state-level simulations. The most striking result in this table is that rejection rates using bootstrap without correction still depend on the size of the treated group. This happens in a dataset with, on average, more than 10,000 observations per group x time cell. In particular, the rejection rate in the simulations with log wages as the outcome variable is zero when the treated group is below the median, and 10% when the treated group is above the median. We present rejection rates using bootstrap with our heteroskedasticity correction in columns 5 and 6. Average rejection rates are around 5%, and we cannot reject that there is no difference in rejection rates above and below the median. However, this test of our method is under-powered, since we estimate rejection rates in the state-level models based on only 408 simulations. In order to provide more precision to estimate the rejection rates of our method, we simulate DID placebo regressions randomly selecting 50 PUMAs in each simulation, which generates many more placebo estimates. These results are presented in panel C of Table ACS. We

[17] Given that we have a limited number of simulations, we do not calculate the average absolute difference in rejection rates across deciles, as we do in the Monte Carlo simulations. For the PUMA-level simulations, there are only approximately 500 simulations for each decile. For the state-level simulations there would be only around 40 simulations for each decile.

also present results DID placebo regressions randomly selecting 25 PUMAs in each simulation in Panel D of Table 6. Remarkably, our method still provides hypothesis testing with correct size regardless of the size of the treated group when $N = 50$ and when $N = 25$.

4.2 Current Population Survey (CPS) Results

We present the simulation results using the CPS in Table 7. Panel A presents rejection rates of DID models using 2 years of data, while Panels B, C and D present rejection rates using respectively 4, 6 and 8 years. Inference with OLS robust standard errors on the individual-level model becomes worse when we include more years of data in the model (column 1). This result is consistent with the findings in Bertrand et al. (2004). The key point is that the panel structure of the CPS Merged Outgoing Rotation Groups generates serial correlation in the errors. We present rejection rates for the residual bootstrap without correction in columns 3 and 4. The average rejection rates are close to 5% irrespective of the number of periods, which was expected given that this method takes serial correlation into account by looking at a linear combination of the residuals (as in CT). However, since this linear combination of the residuals is heteroskedastic, rejection rates based on this method vary with the size of the treated group. We present rejection rates using bootstrap with our heteroskedasticity correction in columns 5 and 6. As in the ACS simulations, we cannot reject that rejection rates have the correct size on average and that rejection rates do not depend on the size of the treated group in all simulations. Therefore, our method is efficient in correcting for heteroskedasticity in a scenario that serial correlation is important without the need to specify the structure of the serial correlation.

4.3 Power with Real Data Simulations

We saw in Sections 4.1 and 4.2 that our method provides tests with correct size in simulations with the ACS and the CPS. We now present in Figure 8 power results from simulations with these datasets. Figure 8.A shows power results using the ACS. When the treated group is above the median, our method is able to detect an effect size of 0.06 log points with probability greater than 70%. When the treated group is below median, we are only able to attain this power for effects greater than 0.1 log points. This again reflects that the variance of $\hat{\alpha}$ is higher when the treated group is smaller. Figures 8.B to 8.E present results for simulations using the CPS with different numbers of time periods. The power in the CPS simulations is considerably lower than in the ACS simulations. The power to reject an effect of 0.06 log points when the treated group is above the median ranges from 26% to 41%, depending on the number of periods used in

the simulations. This happens because the ACS has a much larger number of observations than the CPS. Even though we have only one treated group in all simulations, the larger number of observations in the ACS implies that the group x time variance of the error would be smaller.[18]

As opposed to the power results presented in Section 3.1.2, we do not know the true variance of $\hat{\alpha}$, so it is not possible to compare the power of our method with the power of the UMP test. Still, results from the Monte Carlo simulations suggest that the power of our method should be very close to the power of a UMP test.

5 Conclusion

This paper shows that usual inference methods used in DID models might not perform well in the presence of heteroskedasticity when the number of treated groups is small. In particular, we show that, methods designed to work when there are few treated groups tend to (under-) over-reject the null hypothesis when the number of observations of the treated groups are (large) small relative to the number of observations of the control groups. Using Monte Carlo simulations and simulations with real datasets, we show that this problem is relevant even in datasets with large number of observations per group.

We then provide alternative inference methods that are valid when the number of treated groups is small and there is heteroskedasticity. First, we derive a new inference method that corrects for heteroskedasticity when we use information from the residuals of the control groups to estimate the variance of the treated group. Our method provides asymptotically valid hypothesis testing when the number of control groups goes to infinity even when there is only one treated group. In Monte Carlo simulations and simulations with real datasets, our method provides accurate hypothesis testing with one treated and 24 control groups. We also derive conditions under which an inference method proposed by Abadie et al. (2010) for the Synthetic Control Estimator takes heteroskedasticity into account.

Finally, it is important to point out that our inference method for correcting for heteroskedasticity is more general than the main case we analyzed in this paper, in which the heteroskedasticity is generated by variation in the number of observations per group. In fact, as long as we are able to assume a structure of the variance of a linear combination of the errors, W_j, we are able to apply our method. There are other

[18] For some CPS simulations, the power when the treated group is below median crosses the power when the treated group is above median when the effect size is large. This happens because a large effect size would imply that $\widehat{W}_1^2$ (which is calculated from a model with H_0 imposed) would be large, which would bias our estimate of $G(M)$. Note that this does not invalidate the method, since $\widehat{G(M)}$ is consistent under the null. Also, this distortion only appears when the power of the test was already above 90%.

applications where the variance of W_j might vary by group even when all groups have the same size. This would happen when, for example, Y_{ijt} is a binary variable and average Y_{jt} might be closer or farther away from 0.5 depending on j.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller, "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of Californias Tobacco Control Program," *Journal of the American Statistical Association*, 2010, *105* (490), 493–505.

_ **and Javier Gardeazabal**, "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, March 2003, *93* (1), 113–132.

Angrist, J.D. and J.S. Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, 2009.

Assuncao, J. and B. Ferman, "Does affirmative action enhance or undercut investment incentives? Evidence from quotas in Brazilian Public Universities," *Unpublished Manuscript*, February 2015, *Can be found (as of Feb. 2015), at https://dl.dropboxusercontent.com/u/12654869/Assuncao%20and%20Ferman022015.pdf.*

Bell, R. M. and D. F. McCaffrey, "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples," *Survey Methodology*, 2002, *28* (2), 169–181.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, "How Much Should We Trust Differences-in-Differences Estimates?," *Quarterly Journal of Economics*, 2004, p. 24975.

Brewer, Mike, Thomas F. Crossley, and Robert Joyce, "Inference with Difference-in-Differences Revisited," IZA Discussion Papers 7742, Institute for the Study of Labor (IZA) November 2013.

Cameron, A.C., J.B. Gelbach, and D.L. Miller, "Bootstrap-based improvements for inference with clustered errors," *The Review of Economics and Statistics*, 2008, *90* (3), 414–427.

Canay, Ivan A., Joseph P. Romano, and Azeem M. Shaikh, "Randomization Tests under an Approximate Symmetry Assumption?," 2014.

Conley, Timothy G. and Christopher R. Taber, "Inference with "Difference in Differences with a Small Number of Policy Changes," *The Review of Economics and Statistics*, February 2011, *93* (1), 113–125.

Donald, Stephen G. and Kevin Lang, "Inference with Difference-in-Differences and Other Panel Data," *The Review of Economics and Statistics*, May 2007, *89* (2), 221–233.

Fisher, R.A., *The design of experiments. 1935*, Edinburgh: Oliver and Boyd, 1935.

Hansen, Christian B., "Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects," *Journal of Econometrics*, October 2007, *140* (2), 670–694.

Hausman, Jerry and Guido Kuersteiner, "Difference in difference meets generalized least squares: Higher order properties of hypotheses tests," *Journal of Econometrics*, June 2008, *144* (2), 371–391.

Ibragimov, Rustam and Ulrich K. Mller, "Inference with Few Heterogenous Clusters," 2013.

Imbens, Guido W. and Michal Kolesar, "Robust Standard Errors in Small Samples: Some Practical Advice," Working Paper 18478, National Bureau of Economic Research October 2012.

Lehmann, E.L. and J.P. Romano, *Testing Statistical Hypotheses* Springer Texts in Statistics, Springer New York, 2008.

Liang, KUNG-YEE and SCOTT L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, 1986, *73* (1), 13–22.

MacKinnon, James G. and Matthew D. Webb, "Differences-in-Differences Inference with Few Treated Clusters," 2015.

_ **and** _ , "Wild Bootstrap Inference for Wildly Different Cluster Sizes," Working Papers 1314, Queen's University, Department of Economics February 2015.

Moulton, Brent R., "Random group effects and the precision of regression estimates," *Journal of Econometrics*, August 1986, *32* (3), 385–397.

Webb, Matthew D., "Reworking Wild Bootstrap Based Inference for Clustered Errors," Working Papers 1315, Queen's University, Department of Economics November 2014.

White, Halbert, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, May 1980, *48* (4), 817–838.

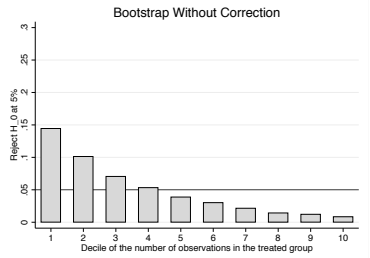
Wooldridge, Jeffrey M., "Cluster-Sample Methods in Applied Econometrics," *American Economic Review*, 2003, *93* (2), 133–138.

Yates, F., "Tests of Significance for 2 2 Contingency Tables," *Journal of the Royal Statistical Society. Series A (General)*, 1984, *147* (3), 426–463.

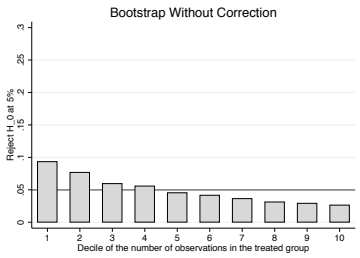
Young, Alwyn, "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results," 2015.

Figure 1: **Rejection Rates in MC Simulations by Decile of M_1, $N = 400$**

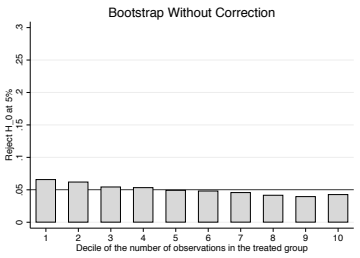
1.A: w/o correction, $\rho = 0.01\%$



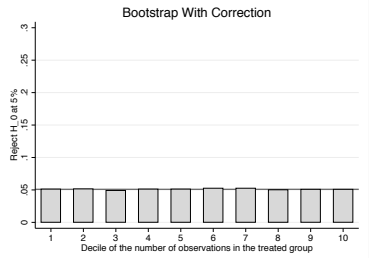
1.B: w/o correction, $\rho = 1\%$



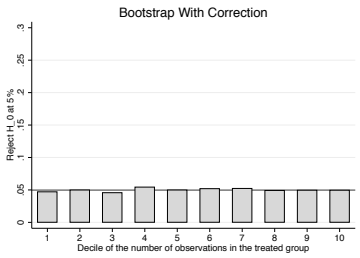
1.C: w/o correction, $\rho = 2\%$



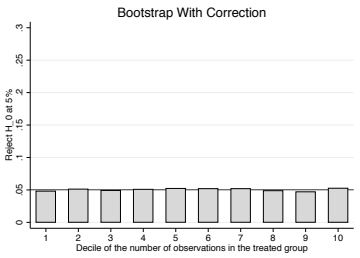
1.D: with correction, $\rho = 0.01\%$



1.E: with correction, $\rho = 1\%$



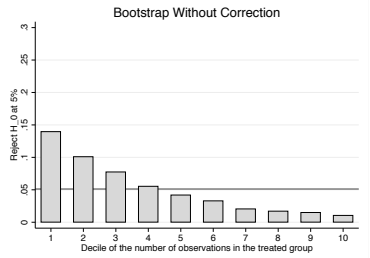
1.F: with correction, $\rho = 2\%$



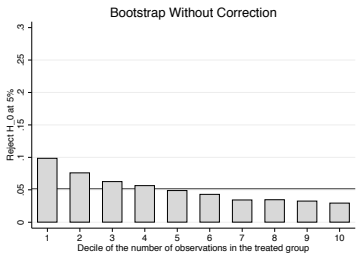
Notes: These figures present the rejection rates conditional on the decile of the number of observation of the treated group when $N = 400$ and $M \in [50, 200]$. These rejection rates are based on Monte Carlos simulations explained in Section 3. Figures 1.A to 1.C present results using the residual bootstrap without correction, while Figures 1.D to 1.F present results using the residual bootstrap method with our heteroskedasticity correction, as explained in Section 2.3.

Figure 2: **Rejection Rates in MC Simulations by Decile of M_1, $N = 100$**

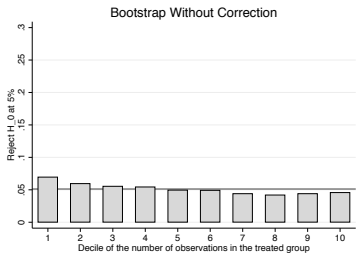
2.A: w/o correction, $\rho = 0.01\%$



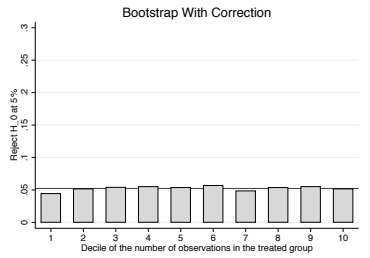
2.B: w/o correction, $\rho = 1\%$



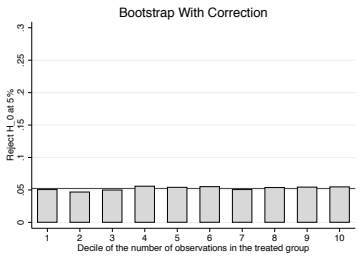
2.C: w/o correction, $\rho = 2\%$



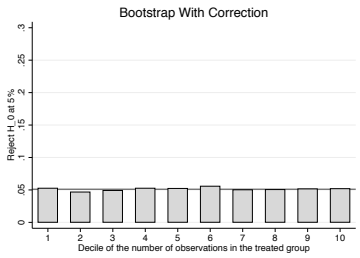
2.D: with correction, $\rho = 0.01\%$



2.E: with correction, $\rho = 1\%$



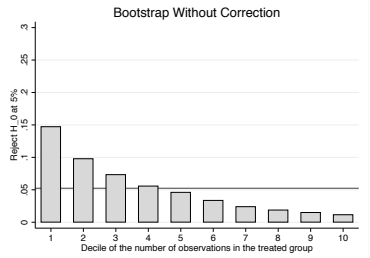
2.F: with correction, $\rho = 2\%$



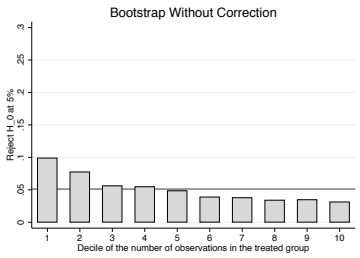
Notes: These figures present the rejection rates conditional on the decile of the number of observation of the treated group when $N = 100$ and $M \in [50, 200]$. These rejection rates are based on Monte Carlos simulations explained in Section 3. Figures 2.A to 2.C present results using the residual bootstrap without correction, while Figures 2.D to 2.F present results using the residual bootstrap method with our heteroskedasticity correction, as explained in Section 2.3.

Figure 3: **Rejection Rates in MC Simulations by Decile of M_1, $N = 50$**

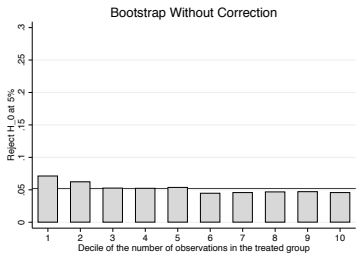
3.A: w/o correction, $\rho = 0.01\%$



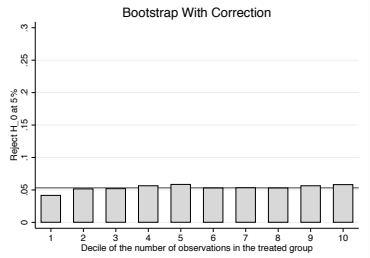
3.B: w/o correction, $\rho = 1\%$



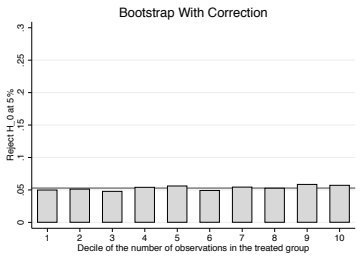
3.C: w/o correction, $\rho = 2\%$



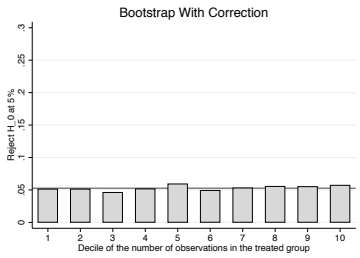
3.D: with correction, $\rho = 0.01\%$



3.E: with correction, $\rho = 1\%$



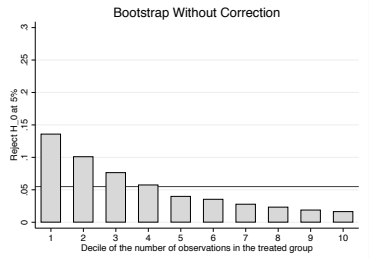
3.F: with correction, $\rho = 2\%$



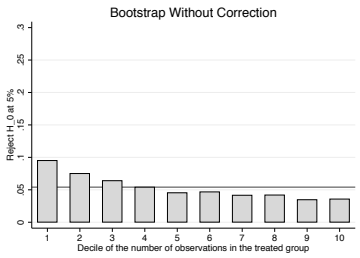
Notes: These figures present the rejection rates conditional on the decile of the number of observation of the treated group when $N = 50$ and $M \in [50, 200]$. These rejection rates are based on Monte Carlos simulations explained in Section 3. Figures 3.A to 3.C present results using the residual bootstrap without correction, while Figures 3.D to 3.F present results using the residual bootstrap method with our heteroskedasticity correction, as explained in Section 2.3.

Figure 4: **Rejection Rates in MC Simulations by Decile of M_1, $N = 25$**

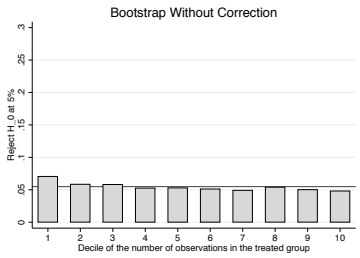
4.A: w/o correction, $\rho = 0.01\%$



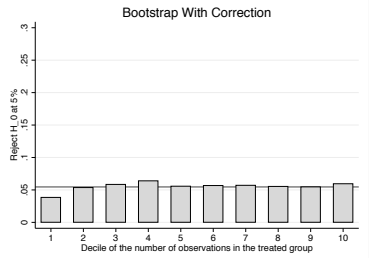
4.B: w/o correction, $\rho = 1\%$



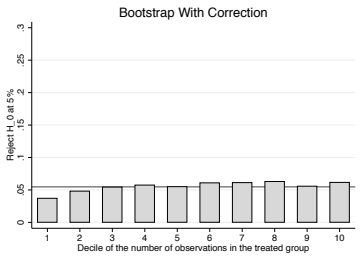
4.C: w/o correction, $\rho = 2\%$



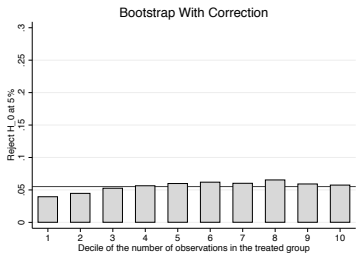
4.D: with correction, $\rho = 0.01\%$



4.E: with correction, $\rho = 1\%$



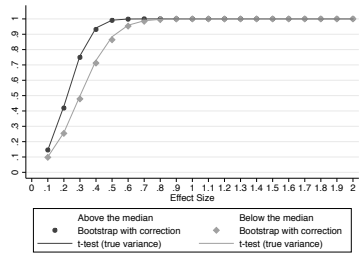
4.F: with correction, $\rho = 2\%$



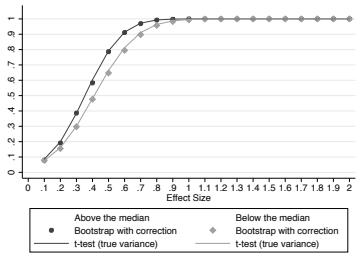
Notes: These figures present the rejection rates conditional on the decile of the number of observation of the treated group when $N = 25$ and $M \in [50, 200]$. These rejection rates are based on Monte Carlos simulations explained in Section 3. Figures 4.A to 4.C present results using the residual bootstrap without correction, while Figures 4.D to 4.F present results using the residual bootstrap method with our heteroskedasticity correction, as explained in Section 2.3.

Figure 5: **Test Power by Treated Group Size - Monte Carlo Simulations with** $N = 400$

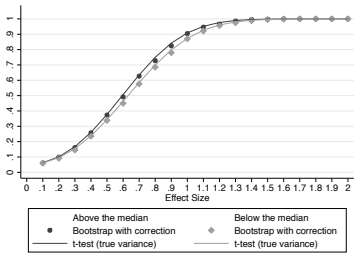
5.A: $M \in [50, 200]$, $\rho = 0.01\%$



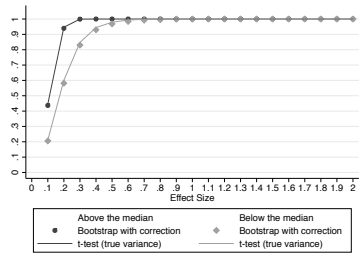
5.B: $M \in [50, 200]$, $\rho = 1\%$



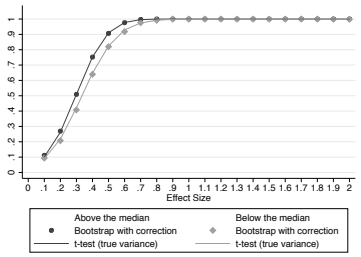
5.C: $M \in [50, 200]$, $\rho = 4\%$



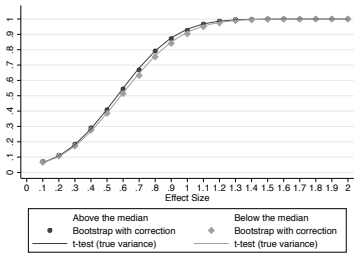
5.D: $M \in [50, 950]$, $\rho = 0.01\%$



5.E: $M \in [50, 950]$, $\rho = 1\%$



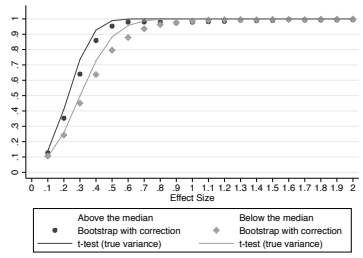
5.F: $M \in [50, 950]$, $\rho = 4\%$



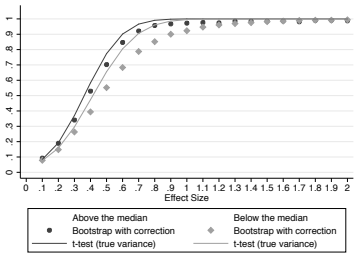
Notes: These figures present the power of the bootstrap with heteroskedasticity correction as a function of the effect size separately when the treated group is above and below the median of group size. The standard deviation of the individual level observation is equal to one across the different scenarios. Therefore, the effect size is in standard deviation terms. Results are based on simulations with total number groups $N = 400$.

Figure 6: **Test Power by Treated Group Size - Monte Carlo Simulations with** $N = 25$

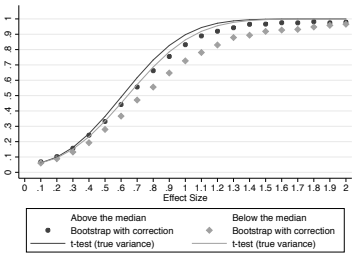
6.A: $M \in [50, 200]$, $\rho = 0.01\%$



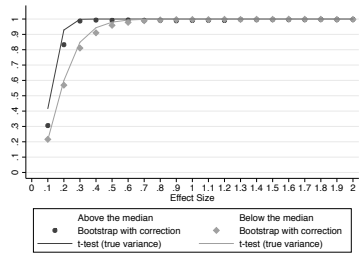
6.B: $M \in [50, 200]$, $\rho = 1\%$



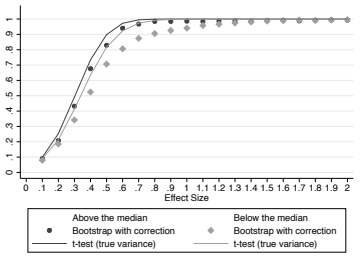
6.C: $M \in [50, 200]$, $\rho = 4\%$



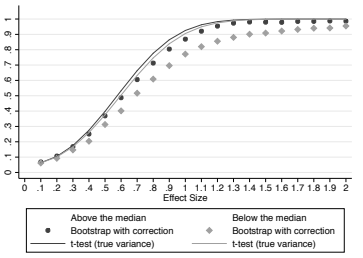
6.D: $M \in [50, 950]$, $\rho = 0.01\%$



6.E: $M \in [50, 950]$, $\rho = 1\%$



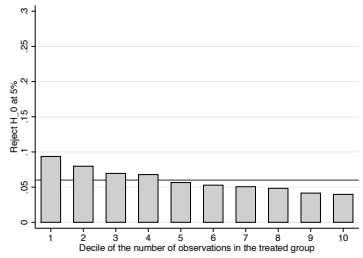
6.F: $M \in [50, 950]$, $\rho = 4\%$



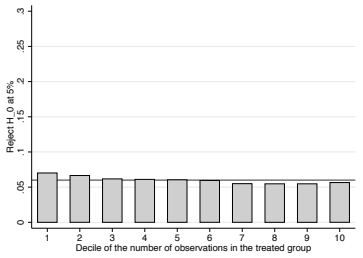
Notes: These figures present the power of the bootstrap with heteroskedasticity correction as a function of the effect size separately when the treated group is above and below the median of group size. The standard deviation of the individual level observation is equal to one across the different scenarios. Therefore, the effect size is in standard deviation terms. Results are based on simulations with total number groups $N = 25$.

Figure 7: **Inference with Synthetic Control**

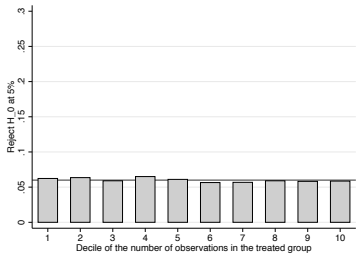
7.A: Just-id, $T = 2$, $\rho = 0.01\%$



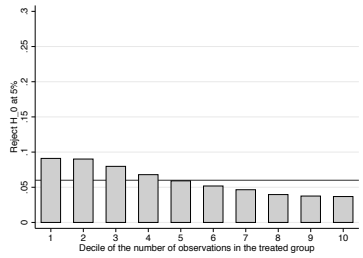
7.B: Just-id, $T = 2$, $\rho = 1\%$



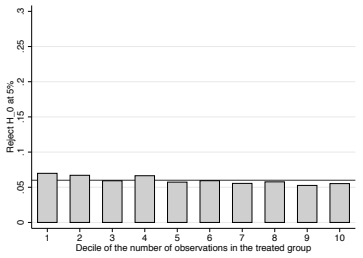
7.C: Just-id, $T = 2$, $\rho = 4\%$



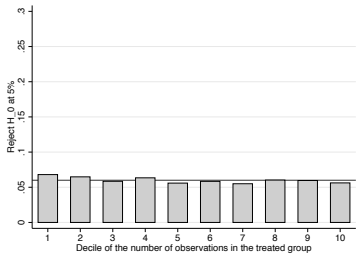
7.D: Just-id, $T = 3$, $\rho = 0.01\%$



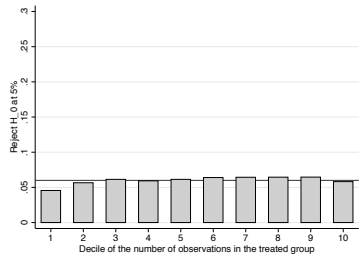
7.E: Just-id, $T = 3$, $\rho = 1\%$



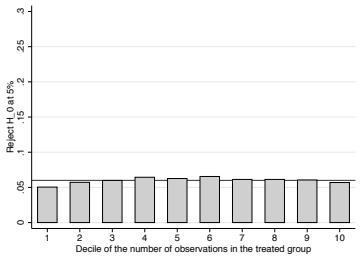
7.F: Just-id, $T = 3$, $\rho = 4\%$



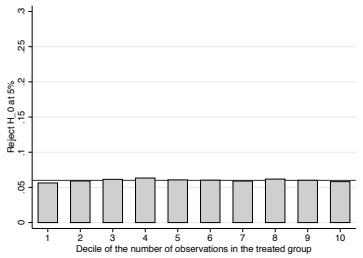
7.G: Over-id, $T = 3$, $\rho = 0.01\%$



7.H: Over-id, $T = 3$, $\rho = 1\%$



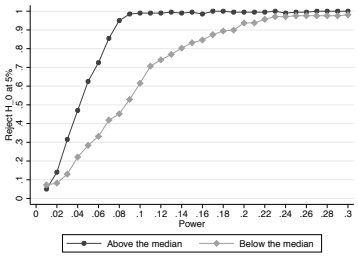
7.I: Over-id, $T = 3$, $\rho = 4\%$



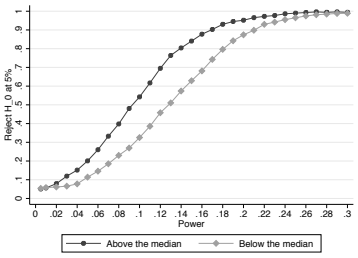
Notes: These figures present rejection rates from Monte Carlo simulations using the inference proposed by Abadie et al. (2010) for the Synthetic Control Estimation for different intra-group correlation parameters (ρ). In all simulations, only one group is treated, $N = 50\%$, and $M \in \{50, 950\}$. Figures 7.A to 7.C report results for a scenario with 2 periods, one pre- and one post-treatment. We estimate the weights using Y_{j1} and M_j. Figures 7.D to 7.F report results for a scenario with 3 periods, two pre- and one post-treatment. We estimate the weights using Y_{j1}, Y_{j2} and M_j. Figures 7.G to 7.I also report results for a scenario with 3 periods, but using only Y_{j1} and M_j to estimate the weights.

Figure 8: **Test Power by Treated Group Size - Simulations with Real Dataset**

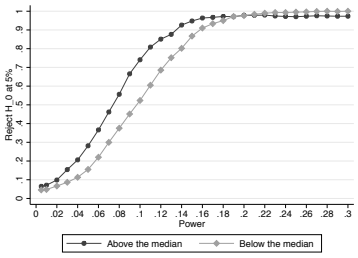
8.A: ACS



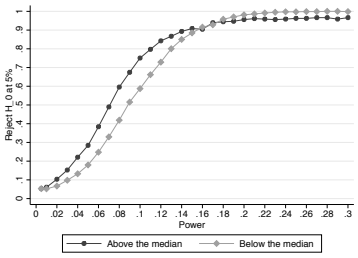
8.B: CPS with $T = 2$



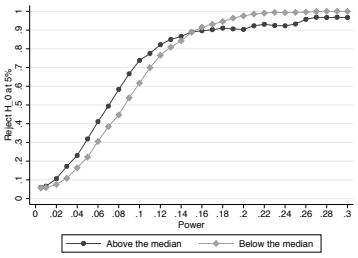
8.C: CPS with $T = 4$



8.D: CPS with $T = 6$



8.E: CPS with $T = 8$



Notes: These figures present the power of the bootstrap with heteroskedasticity correction for simulations using real datasets. Results are presented separately when the treated group is above and below the median of group size. The outcome variable is log wages, and effect sizes are measured in log points. Figure 8.A presents results using the ACS, while Figures 8.B to 8.E present results using the CPS with varying number of periods.

Table 1: **Rejection Rates in MC Simulations with $N_0 + 1 = 400$**

	Inference Method									
	Robust OLS		Donald and Lang		Conley and Taber		Bootstrap w/o correction		Bootstrap with correction	
ρ	Mean (1)	Absolute Difference (2)	Mean (3)	Absolute Difference (4)	Mean (5)	Absolute Difference (6)	Mean (7)	Absolute Difference (8)	Mean (9)	Absolute Difference (10)
					Panel A: $M \in [50, 200]$					
0.01%	0.054	0.002	0.053	0.039	0.050	0.036	0.049	0.034	0.051	0.001
1%	0.192	0.036	0.050	0.019	0.050	0.018	0.049	0.017	0.050	0.002
4%	0.420	0.059	0.049	0.007	0.050	0.006	0.050	0.007	0.050	0.002
					Panel B: $M \in [200, 800]$					
0.01%	0.057	0.002	0.053	0.036	0.051	0.034	0.049	0.034	0.049	0.002
1%	0.415	0.065	0.051	0.008	0.049	0.006	0.050	0.008	0.050	0.002
4%	0.661	0.051	0.049	0.004	0.051	0.003	0.050	0.003	0.050	0.002
					Panel C: $M \in [50, 950]$					
0.01%	0.057	0.003	0.054	0.061	0.051	0.057	0.050	0.057	0.051	0.002
1%	0.396	0.098	0.051	0.019	0.051	0.019	0.049	0.018	0.050	0.001
4%	0.637	0.093	0.051	0.006	0.049	0.006	0.050	0.006	0.049	0.002

Note: This table presents results from Monte Carlo simulations with 400 groups, as explained in Section 3. In all simulations, only one group is treated. Each line presents simulation for different values of intra-group correlation, while each panel presents results for different numbers of observations per group. We consider 5 inference methods: hypothesis testing using robust standard errors from the individual level regression, DL, CT, cluster residual bootstrap without correction, and cluster residual bootstrap with our heteroskedasticity correction. For the bootstrap methods, we imposed H_0, and we used the wild bootstrap idea of randomizing whether we multiply the residuals by 1 or -1. For each inference method, we report the average rejection rate for a 5% significance level test. We also report a measure of how rejection rates depend on the number of observations in the treated group, which we call "absolute difference". To construct this measure, we calculate the absolute difference in rejection rates for each decile of M_1 relative to the average rejection rate, and then we average these absolute differences across deciles. We run 100,000 simulations for each $M \times \rho \times N_0$ scenario. The standard error for the average rejection rates is around 0.07 percentage points, while the standard error for the absolute difference is around 0.04-0.07 percentage points.

Table 2: **Rejection Rates in MC Simulations with $N_0 + 1 = 100$**

	Inference Method									
	Robust OLS		Donald and Lang		Conley and Taber		Bootstrap w/o correction		Bootstrap with correction	
ρ	Mean (1)	Absolute Difference (2)	Mean (3)	Absolute Difference (4)	Mean (5)	Absolute Difference (6)	Mean (7)	Absolute Difference (8)	Mean (9)	Absolute Difference (10)
					Panel A: $M \in [50, 200]$					
0.01%	0.054	0.003	0.054	0.036	0.049	0.032	0.051	0.034	0.052	0.003
1%	0.193	0.032	0.052	0.017	0.049	0.018	0.052	0.017	0.052	0.002
4%	0.418	0.062	0.052	0.008	0.047	0.007	0.051	0.007	0.051	0.002
					Panel B: $M \in [200, 800]$					
0.01%	0.057	0.001	0.052	0.037	0.049	0.032	0.050	0.033	0.050	0.002
1%	0.415	0.058	0.050	0.008	0.049	0.008	0.052	0.007	0.052	0.002
4%	0.658	0.049	0.050	0.004	0.048	0.003	0.052	0.002	0.053	0.002
					Panel C: $M \in [50, 950]$					
0.01%	0.057	0.002	0.057	0.060	0.049	0.053	0.050	0.054	0.052	0.003
1%	0.400	0.095	0.050	0.019	0.049	0.018	0.050	0.017	0.051	0.002
4%	0.636	0.089	0.049	0.006	0.048	0.005	0.052	0.006	0.051	0.001

Note: This table presents results from Monte Carlo simulations with 100 groups, as explained in Section 3. In all simulations, only one group is treated. Each line presents simulation for different values of intra-group correlation, while each panel presents results for different numbers of observations per group. We consider 5 inference methods: hypothesis testing using robust standard errors from the individual level regression, DL, CT, cluster residual bootstrap without correction, and cluster residual bootstrap with our heteroskedasticity correction. For the bootstrap methods, we imposed H_0, and we used the wild bootstrap idea of randomizing whether we multiply the residuals by 1 or -1. For each inference method, we report the average rejection rate for a 5% significance level test. We also report a measure of how rejection rates depend on the number of observations in the treated group, which we call "absolute difference". To construct this measure, we calculate the absolute difference in rejection rates for each decile of M_1 relative to the average rejection rate, and then we average these absolute differences across deciles. We run 100,000 simulations for each $M \times \rho \times N_0$ scenario. The standard error for the average rejection rates is around 0.07 percentage points, while the standard error for the absolute difference is around 0.04-0.07 percentage points.

Table 3: **Rejection Rates in MC Simulations with $N_0 + 1 = 50$**

	Inference Method									
	Robust OLS		Donald and Lang		Conley and Taber		Bootstrap w/o correction		Bootstrap with correction	
ρ	Mean	Absolute Difference	Mean	Absolute Difference	Mean	Absolute Difference	Mean	Absolute Difference	Mean	Absolute Difference
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
					Panel A: $M \in [50, 200]$					
0.01%	0.052	0.003	0.054	0.035	0.046	0.030	0.052	0.033	0.053	0.003
1%	0.192	0.037	0.051	0.017	0.046	0.014	0.051	0.016	0.053	0.003
4%	0.420	0.057	0.050	0.006	0.045	0.005	0.052	0.006	0.053	0.003
					Panel B: $M \in [200, 800]$					
0.01%	0.057	0.002	0.053	0.034	0.047	0.029	0.051	0.031	0.052	0.003
1%	0.415	0.060	0.049	0.007	0.047	0.006	0.052	0.006	0.052	0.003
4%	0.663	0.047	0.049	0.002	0.047	0.002	0.051	0.002	0.052	0.003
					Panel C: $M \in [50, 950]$					
0.01%	0.056	0.002	0.057	0.060	0.046	0.048	0.050	0.052	0.052	0.004
1%	0.398	0.099	0.051	0.019	0.047	0.017	0.051	0.015	0.051	0.004
4%	0.635	0.089	0.050	0.006	0.046	0.005	0.051	0.003	0.051	0.005

Note: This table presents results from Monte Carlo simulations with 50 groups, as explained in Section 3. In all simulations, only one group is treated. Each line presents simulation for different values of intra-group correlation, while each panel presents results for different numbers of observations per group. We consider 5 inference methods: hypothesis testing using robust standard errors from the individual level regression, DL, CT, cluster residual bootstrap without correction, and cluster residual bootstrap with our heteroskedasticity correction. For the bootstrap methods, we imposed H_0, and we used the wild bootstrap idea of randomizing whether we multiply the residuals by 1 or -1. For each inference method, we report the average rejection rate for a 5% significance level test. We also report a measure of how rejection rates depend on the number of observations in the treated group, which we call "absolute difference". To construct this measure, we calculate the absolute difference in rejection rates for each decile of M_1 relative to the average rejection rate, and then we average these absolute differences across deciles. We run 100,000 simulations for each $M \times \rho \times N_0$ scenario. The standard error for the average rejection rates is around 0.07 percentage points, while the standard error for the absolute difference is around 0.04-0.07 percentage points.

Table 4: **Rejection Rates in MC Simulations with $N_0 + 1 = 25$**

	Inference Method									
	Robust OLS		Donald and Lang		Conley and Taber		Bootstrap w/o correction		Bootstrap with correction	
ρ	Mean (1)	Absolute Difference (2)	Mean (3)	Absolute Difference (4)	Mean (5)	Absolute Difference (6)	Mean (7)	Absolute Difference (8)	Mean (9)	Absolute Difference (10)
					Panel A: $M \in [50, 200]$					
0.01%	0.052	0.002	0.053	0.033	0.078	0.038	0.053	0.032	0.055	0.004
1%	0.193	0.032	0.051	0.016	0.079	0.020	0.053	0.015	0.055	0.005
4%	0.424	0.055	0.050	0.006	0.079	0.008	0.054	0.005	0.056	0.006
					Panel B: $M \in [200, 800]$					
0.01%	0.056	0.002	0.053	0.031	0.077	0.037	0.051	0.029	0.056	0.006
1%	0.417	0.060	0.049	0.009	0.078	0.008	0.054	0.005	0.056	0.006
4%	0.664	0.048	0.050	0.005	0.079	0.003	0.055	0.001	0.054	0.007
					Panel C: $M \in [50, 950]$					
0.01%	0.057	0.003	0.056	0.055	0.076	0.059	0.047	0.045	0.056	0.004
1%	0.403	0.091	0.052	0.015	0.077	0.019	0.052	0.015	0.056	0.007
4%	0.643	0.084	0.052	0.004	0.080	0.006	0.054	0.005	0.055	0.007

Note: This table presents results from Monte Carlo simulations with 25 groups, as explained in Section 3. In all simulations, only one group is treated. Each line presents simulation for different values of intra-group correlation, while each panel presents results for different numbers of observations per group. We consider 5 inference methods: hypothesis testing using robust standard errors from the individual level regression, DL, CT, cluster residual bootstrap without correction, and cluster residual bootstrap with our heteroskedasticity correction. For the bootstrap methods, we imposed H_0, and we used the wild bootstrap idea of randomizing whether we multiply the residuals by 1 or -1. For each inference method, we report the average rejection rate for a 5% significance level test. We also report a measure of how rejection rates depend on the number of observations in the treated group, which we call "absolute difference". To construct this measure, we calculate the absolute difference in rejection rates for each decile of M_1 relative to the average rejection rate, and then we average these absolute differences across deciles. We run 100,000 simulations for each $M \times \rho \times N_0$ scenario. The standard error for the average rejection rates is around 0.07 percentage points, while the standard error for the absolute difference is around 0.04-0.07 percentage points.

Table 5: **Number of Observations per Group x Time cell**

	ACS		CPS
	PUMA	State	State
	(1)	(2)	(3)
Average	778.12	10,137.79	771.23
1%	129	883	119
5%	157	1,037	355
10%	174	1,250	392
25%	218	2,527	464
50%	338	7,205	546
75%	703	11,509	775
90%	1,418	21,099	1,709
95%	2,469	32,961	1,937
99%	9,555	62,752	3,297

Note: This Table presents the distribution of number of observations per groups in the simulations with real datasets (Section 4). Column 1 presents information for PUMA-level ACS simulations, column 2 presents information for state-level ACS simulations, while column 3 presents information for state-level CPS simulations.

Table 6: **Simulations with the ACS Survey**

	Inference Method					
	Robust OLS		Bootstrap w/o correction		Bootstrap with correction	
Outcome Variable	Mean (1)	Diff (2)	Mean (3)	Diff (4)	Mean (5)	Diff (6)
Panel A: ACS with PUMA-level interventions						
Employment	0.072***	0.010	0.048	-0.080***	0.050	0.000
	(0.004)	(0.008)	(0.003)	(0.006)	(0.003)	(0.006)
Log(wages)	0.081***	0.000	0.048	-0.080***	0.050	0.005
	(0.004)	(0.008)	(0.003)	(0.006)	(0.003)	(0.006)
Panel B: ACS with state-level interventions						
Employment	0.064	0.003	0.044	-0.087***	0.051	-0.013
	(0.011)	(0.021)	(0.010)	(0.020)	(0.011)	(0.022)
Log(wages)	0.081**	-0.021	0.051	-0.101***	0.054	-0.027
	(0.015)	(0.031)	(0.011)	(0.021)	(0.011)	(0.022)
Panel C: ACS with PUMA-level interventions, $N = 50$						
Employment	0.072***	0.001	0.043***	-0.069***	0.051	-0.004
	(0.004)	(0.007)	(0.002)	(0.005)	(0.003)	(0.005)
Log(wages)	0.084***	-0.001	0.045**	-0.070***	0.051	0.001
	(0.004)	(0.008)	(0.002)	(0.005)	(0.003)	(0.005)
Panel D: ACS with PUMA-level interventions, $N = 25$						
Employment	0.069***	0.009	0.040***	-0.057***	0.050	-0.001
	(0.004)	(0.007)	(0.002)	(0.004)	(0.003)	(0.005)
Log(wages)	0.082***	0.000	0.039***	-0.059***	0.050	-0.004
	(0.004)	(0.008)	(0.002)	(0.004)	(0.003)	(0.005)

Note: This table presents rejection rates for the simulations using ACS data. For each pair of consecutive years, we run a DID regression using one group as treated and the other groups as a control. The outcome variable is employment status or log(wages) for women aged between 25 and 40. Then we test the hypothesis that the effect of the "intervention" is equal to zero using different inference methods: hypothesis testing using robust standard errors from individual level DID model, bootstrap without and bootstrap with our heteroskedasticity correction. Panel A reports results when groups are defined as PUMAs, while Panel B reports results when groups are defined as states. In Panels C and D we present results with PUMA-level treatments using 50 and 25 randomly selected PUMAs. We report average rejection rate and the difference in rejection rates when the size of the treated group is above or below the median. Given that we have a limited number of simulations, we do not calculate the average absolute difference in rejection rates across deciles, as we do in the Monte Carlo simulations. We present in brackets standard errors for the rejection rates. For Panels C and D, standard errors are clustered at the treated group x year level. For average rejection rates (columns 1, 3, and 5), * means that we reject at 10% that the average rejection rate is equal to 5%, while for the differences in rejection rates (columns 2, 4, and 6) * means that we reject at 10% that rejection rate for M_1 above and below the median are equal. ** means that we reject at 5%, while *** means that we reject at 1%.

Table 7: **Simulations with the CPS Survey**

	Inference Method					
	Robust OLS		Bootstrap w/o correction		Bootstrap with correction	
Outcome Variable	Mean (1)	Diff (2)	Mean (3)	Diff (4)	Mean (5)	Diff (6)
			Panel A: 2 years			
Employment	0.047	-0.003	0.046	-0.041***	0.050	0.002
	(0.005)	(0.010)	(0.005)	(0.010)	(0.005)	(0.010)
Log(wages)	0.066***	-0.011	0.045	-0.047***	0.051	0.003
	(0.006)	(0.012)	(0.005)	(0.010)	(0.005)	(0.010)
			Panel B: 4 years			
Employment	0.062**	0.016	0.043	-0.041***	0.053	-0.016
	(0.006)	(0.012)	(0.005)	(0.010)	(0.005)	(0.011)
Log(wages)	0.102***	0.024	0.048	-0.042***	0.053	0.008
	(0.007)	(0.016)	(0.005)	(0.010)	(0.005)	(0.011)
			Panel C: 6 years			
Employment	0.087***	0.001	0.052	-0.050***	0.052	-0.016
	(0.007)	(0.015)	(0.006)	(0.011)	(0.006)	(0.011)
Log(wages)	0.143***	0.059***	0.050	-0.045***	0.051	-0.008
	(0.009)	(0.019)	(0.005)	(0.011)	(0.006)	(0.011)
			Panel C: 8 years			
Employment	0.135***	0.044**	0.043	-0.040***	0.045	-0.010
	(0.009)	(0.020)	(0.005)	(0.010)	(0.005)	(0.011)
Log(wages)	0.207***	0.043*	0.045	-0.029***	0.049	0.005
	(0.011)	(0.023)	(0.005)	(0.011)	(0.006)	(0.011)

Note: This table presents rejection rates for the simulations using CPS data. In each simulation, we run a DID regression using one group as treated and the other groups as a control. The outcome variable is employment status or log(wages) for women aged between 25 and 40. Then we test the hypothesis that the effect of the "intervention" is equal to zero using different inference methods: hypothesis testing using robust standard errors from individual level DID model, bootstrap without and bootstrap with our heteroskedasticity correction. Panel A reports results of DID models using 2 consecutive years of data, while Panels B and C report results of DID models using respectively 4 and 6 consecutive years of data. We report average rejection rate and the difference in rejection rates when the size of the treated group is above or below the median. Given that we have a limited number of simulations, we do not calculate the average absolute difference in rejection rates across deciles, as we do in the Monte Carlo simulations. We present in brackets standard errors for the rejection rates. For Panels C and D, standard errors are clustered at the treated group x year level. For average rejection rates (columns 1, 3, and 5), * means that we reject at 10% that the average rejection rate is equal to 5%, while for the differences in rejection rates (columns 2, 4, and 6) * means that we reject at 10% that rejection rate for M_1 above and below the median are equal. ** means that we reject at 5%, while *** means that we reject at 1%.

Supplemental Appendix: Inference in Differences-in-Differences with Different Group Sizes

This supplemental appendix contains the main theorems and proofs of the paper "Inference in Differences-in-Differences with Different Group Sizes". We use the same notation as in the main paper. Let $M(j,t)$ be the number of observations in group j, time t.

The aggregated model is:

$$y_{jt} = \alpha d_{jt} + \theta_j + \gamma_t + \eta_{jt} \tag{14}$$

We assume T periods of time ($t = 1, .., N$) and N_1 treated groups and N_0 control groups in such a way that $N_0 + N_1 = N$. Consider the restricted model in which we impose the null hypothesis, $H_0 : \alpha = \alpha_0$,

$$y_{jt} = \alpha_0 d_{jt} + \theta_j + \gamma_t + \eta_{jt}$$

We will work with a linear combination of the residuals of this regression,

$$\widehat{W}_j^R = \frac{1}{T - t^*} \sum_{t=t^*+1}^{T} \widehat{\eta}_{jt}^R - \frac{1}{t^*} \sum_{t=1}^{t^*} \widehat{\eta}_{jt}^R$$

We can calculate the DID coefficient $\widehat{\alpha}$ based on a linear combination of $\widehat{W}_j^R$. Define the operator $\bigtriangledown Y_j = \frac{1}{T-t^*} \sum_{t=t^*+1}^{T} Y_{jt} - \frac{1}{t^*} \sum_{t=1}^{t^*} Y_{jt}$. We can write $\widehat{\alpha}$ as:

$$\widehat{\alpha} = \frac{1}{N_1} \sum_{j=1}^{N_1} \bigtriangledown Y_1 - \frac{1}{N_0} \sum_{j=N_1+1}^{N} \bigtriangledown Y_j$$

Since $\widehat{y}_{jt} = \alpha_0 d_{jt} + \widehat{\theta}_j + \widehat{\gamma}_t$, then $\bigtriangledown \widehat{Y}_j^R = \alpha_0 + \frac{1}{T-t^*} \sum_{t=t^*+1}^{T} \widehat{\gamma}_t - \frac{1}{t^*} \sum_{t=1}^{t^*} \widehat{\gamma}_t$ for $j = 1, ..., N_1$ and $\bigtriangledown \widehat{Y}_j^R = \frac{1}{T-t^*} \sum_{t=t^*+1}^{T} \widehat{\gamma}_t - \frac{1}{t^*} \sum_{t=1}^{t^*} \widehat{\gamma}_t$ for $j = N_1 + 1, ..., N$.

Therefore:

$$\widehat{\alpha} - \alpha_0 = \frac{1}{N_1} \sum_{j=1}^{N_1} \widehat{W}_j^R - \frac{1}{N_0} \sum_{j=N_1+1}^{N} \widehat{W}_j^R$$

We define W_j as a linear combination of the error terms,

$$W_j^R = \frac{1}{T - t^*} \sum_{t=t^*+1}^{T} \eta_{jt}^R - \frac{1}{t^*} \sum_{t=1}^{t^*} \eta_{jt}^R$$

We impose assumptions about the behavior of W_j. We assume that T is fixed.

Assumption 1 (Independence and Distribution): W_j's are independent across j and the distribution of W_j only differs among the j by the variance.

Assumption 2[19] **(Exogeneity and Variance-Covariance Structure)**:

$$E\left[W_j^R\right] = 0$$

[19]This assumption can be derived from assumptions about η_{jt} or about the unobservable terms in the individual-level model. However, this assumption is general, allowing serial correlation of the η_{jt}.

$$Var\left[W_j^R\right] = A + \widetilde{B}\left(\frac{1}{(T-t^*)^2}\sum_{t=t^*+1}^{T}\frac{1}{M(j,t)} + \frac{1}{(t^*)^2}\sum_{t=1}^{t^*}\frac{1}{M(j,t)}\right)$$
$$= A + \widetilde{B}\cdot h(M(j,t))$$

where A and $\widetilde{B}$ are constants, and $h(M(j,t)) \equiv \frac{1}{(T-t^*)^2}\sum_{t=t^*+1}^{T}\frac{1}{M(j,t)} + \frac{1}{(t^*)^2}\sum_{t=1}^{t^*}\frac{1}{M(j,t)}$. For simplicity, in the paper, we work with the case in which $M(j,t) = M_j$. In this case, the variance simplifies to

$$Var\left[W_j^R\right] = A + \frac{B}{M_j}$$

for a constant B.

Under assumptions 1 and 2, the variance of this DID estimator is

$$Var\left[\widehat{\alpha} - \alpha_0\right] = A\left(\frac{N_1+N_0}{N_1N_0}\right) + \widetilde{B}\left(\frac{1}{N_1^2}\sum_{j=1}^{N_1}h(M(j,t)) + \frac{1}{N_0^2}\sum_{j=N_1+1}^{N}h(M(j,t))\right) \tag{15}$$

We assume that the number of individuals in each group is fixed and does not vary with N_0. As $N_0 \to \infty$,

$$\widehat{\alpha} - \alpha_0 \to \frac{1}{N_1}\sum_{j=1}^{N_1}\widehat{W}_j^R$$

$$Var\left[\widehat{\alpha}\right] \to \frac{A}{N_1} + \widetilde{B}\left(\frac{1}{N_1^2}\sum_{j=1}^{N_1}h(M(j,t))\right)$$

If we know the variance of W_j^R, we could re-scale the residuals $\widehat{W}_j^R$ and use a cluster residual bootstrap on the re-scaled residuals even if the model is heteroskedastic. The idea is to normalize $\widehat{W}_j^R$ such that $\widehat{W}_j^{norm} = \widehat{W}_j^R \cdot \sqrt{\frac{1}{Var\left[W_j^R\right]}}$, and then generate a bootstrap sample using the re-scaled residuals $\widetilde{\widehat{W}}_{j,b} = \widehat{W}_{j,b}^{norm} \cdot \sqrt{Var\left[W_j^R\right]}$, and use the residuals $\widetilde{\widehat{W}}_{j,b}$ to estimate $\widehat{\alpha}_b - \alpha_0$,

$$\widehat{\alpha}_b - \alpha_0 = \frac{1}{N_1}\sum_{j=1}^{N_1}\widetilde{\widehat{W}}_{j,b} - \frac{1}{N_0}\sum_{j=N_1+1}^{N}\widetilde{\widehat{W}}_{j,b}$$

where b indicates each re-sampling, $b = 1, ..., \mathcal{B}$. In each re-sampling, we calculate $\widehat{\alpha}_b$. We reject H_0 at level α if and only if $\widehat{\alpha} - \alpha_0 < (\widehat{\alpha}_b - \alpha_0)[\frac{\alpha}{2}]$ or $\widehat{\alpha} - \alpha_0 > (\widehat{\alpha}_b - \alpha_0)[1 - \frac{\alpha}{2}]$, where $(\widehat{\alpha}_b - \alpha_0)[q]$ denotes the qth quantile of the distribution of $\{(\widehat{\alpha}_1 - \alpha_0), ..., (\widehat{\alpha}_{\mathcal{B}} - \alpha_0)\}$.

Theorem 1 *Define $d^*_{1-\frac{\alpha}{2}}$ and $d^*_{\frac{\alpha}{2}}$ as the $(1-\frac{\alpha}{2})$th and $\frac{\alpha}{2}$th quantile of the empirical distribution of $(\widehat{\alpha}_b - \alpha_0)$, $b = 1, ..., \mathcal{B}$. Assuming that we know the variance of W_j^R, under assumptions 1 and 2,*

$$\Pr\left[d^*_{1-\frac{\alpha}{2}} \le \widehat{\alpha} - \alpha_0 \le d^*_{\frac{\alpha}{2}}\middle|\alpha_0\right] \to_p 1-\alpha$$

Proof. We divide this proof in two parts. Define $\Gamma_j(w) \equiv \Pr\left[\sum_{j=1}^{N_1}W_j^R < w\right]$ and $\widehat{\Gamma}_{j,b}(w) = \Pr\left[\sum_{j=1}^{N_1}\widehat{W}_{j,b}^R < w\right]$. First we show that $\widehat{\Gamma}_{j,b}(w)$ converges in probability to $\Gamma_j(w)$ uniformly on any compact subset of the support of W, as $N_0 \to \infty$ and $\mathcal{B} \to \infty$. Then, we show that $\Pr\left[d^*_{1-\frac{\alpha}{2}} \le \widehat{\alpha} - \alpha_0 \le d^*_{\frac{\alpha}{2}}\middle|\alpha_0\right] \to_p 1-\alpha$.

Since under our assumptions, $W_j^{'}$s are independent across j and have the same distribution except by the variance, we can

write

$$\begin{aligned}\Gamma_j(w) &= \Pr\left[\sum_{j=1}^{N_1} W_j^R < w\right] \\ &= \int ... \int 1\left\{\sum_{j=1}^{N_1} W_j^R < w\right\} dF_1\left(W_1^R\right) \cdot dF_2\left(W_2^R\right) \cdot ... \cdot dF_{N_1}\left(W_{N_1}^R\right)\end{aligned}$$

and

$$\begin{aligned}\widehat{\Gamma}_{j,b}(w) &= \Pr\left[\sum_{j=1}^{N_1} \widehat{W}_{j,b}^R < w_{j,b}\right] \\ &= \int ... \int 1\left\{\sum_{j=1}^{N_1} \widehat{W}_j^R < w\right\} d\widehat{F}_1\left(W_1^R\right) \cdot d\widehat{F}_2\left(W_2^R\right) \cdot ... \cdot d\widehat{F}_{N_1}\left(W_{N_1}^R\right)\end{aligned}$$

In order to estimate this distribution, we use $\widehat{F}_j(.)$ which is the empirical CDF obtained using the re-scaled residuals $\widetilde{W}_{j,b} = \widehat{W}_{j,b}^R \cdot \sqrt{\frac{Var\left[W_j^R\right]}{Var\left[W_b^R\right]}}$,

$$\begin{aligned}\widehat{F}_{j,b}\left(w_{j,b}\right) &= \frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}} 1\{\widetilde{W}_{j,b} < w_{j,b}\} \\ &= \frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}} 1\left\{\widehat{W}_{j,b}^R \cdot \sqrt{\frac{Var\left[W_j^R\right]}{Var\left[W_{j,b}^R\right]}} < w_j \cdot \sqrt{\frac{Var\left[W_j^R\right]}{Var\left[W_{j,b}^R\right]}}\right\}\end{aligned}$$

where $w_{j,b} = w_j \cdot c_{jb}$, with $c_{jb} = \sqrt{\frac{Var\left[W_j^R\right]}{Var\left[W_{j,b}^R\right]}}$ .In this case, c_{jb} is a constant.

Define $\widehat{F}_{j,b}^*\left(w_{j,bj}\right) = \frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}} 1\{W_{j,b}^R < w_{j,b}\}$. Note that

$$\begin{aligned}\sup_{w_j \in \Theta}\left|\widehat{F}_{j,b}\left(w_{j,b}\right) - \Gamma_j(w)\right| &= \sup_{w_j \in \Theta}\left|\widehat{F}_{j,b}\left(w_{j,b}\right) - \widehat{F}_{j,b}^*\left(w_{j,b}\right) + \widehat{F}_{j,b}^*\left(w_{j,b}\right) - F_j(w)\right| \\ &\leq \sup_{w_j \in \Theta}\left|\widehat{F}_{j,b}\left(w_{j,b}\right) - \widehat{F}_{j,b}^*\left(w_{j,b}\right)\right| + \sup_{w_j \in \Theta}\left|\widehat{F}_{j,b}^*\left(w_{j,b}\right) - F_j(w)\right|\end{aligned}$$

Define ι_T as a vector Tx1 of $1's$ and ι_N as a vector Tx1 of $1's$. and note that,

$$\begin{aligned}\widehat{\eta}_{jt}^R &= y_{jt} - \widehat{\theta}_j - \widehat{\gamma}_t \\ &= \widetilde{\widetilde{y}}_{jt} = \widetilde{\widetilde{\eta}}_{jt}\end{aligned}$$

where $\widetilde{\widetilde{y}}_{jt} = (1 - P_T)(1 - P_N)y_{jt}$ and $\widetilde{\widetilde{\eta}}_{jt} = (1 - P_T)(1 - P_N)\eta_{jt}$, where $P_T = \iota_T\left(\iota_T'\iota_T\right)^{-1}\iota_T'$ and $P_N = \iota_N\left(\iota_N'\iota_N\right)^{-1}\iota_N'$. As $N_0 \to \infty$, $\widetilde{\widetilde{\eta}}_{jt} \to (1 - P_T)\eta_{jt}$, and we can show that

$$\widehat{W}_j^R \to \frac{1}{T - t^*}\sum_{t=t^*+1}^{T} \eta_{jt} - \frac{1}{t^*}\sum_{t=1}^{t^*} \eta_{jt}$$

$$\begin{aligned}\sup_{w_j\in\Theta}\left|\widehat{F}_{j,b}\left(w_{j,b}\right)-\widehat{F}^{*}_{j,b}\left(w_{j,b}\right)\right| &= \sup_{w_j\in\Theta}\left|\frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}}\left(1\{\widehat{W}^{R}_{j,b}<w_{j,b}\}-1\{W^{R}_{j,b}<w_{j,b}\}\right)\right| \\ &\leq \frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}}\sup_{w_j\in\Theta}\left|1\{\widehat{W}^{R}_{j,b}<w_{j,b}\}-1\{W^{R}_{j,b}<w_{j,b}\}\right| \\ &= o\left(1\right)\end{aligned}$$

Now, we work with the second term.

$$\begin{aligned}\sup_{w_j\in\Theta}\left|\widehat{F}^{*}_{j,b}\left(w_{j,b}\right)-\Gamma_j\left(w\right)\right| \leq \sup_{w_j\in\Theta}&\left|\widehat{F}^{*}_{j,b}\left(w_{j,b}\right)-F_{j,b}\left(w_{j,b}\right)\right|+ \\ &\sup_{w_j\in\Theta}\left|F_{j,b}\left(w_{j,b}\right)-F_j\left(w\right)\right|\end{aligned}$$

where $F_{j,b}\left(w_{j,b}\right)$ is the cumulative distribution function of $W_{j,b}$. Note that $W_{j,b}$ are independent across j, have the same distribution and the same variance that equals de variance of W_j. By the Glivenko-Cantelli Theorem,

$$\sup_{w_j\in\Theta}\left|\widehat{F}^{*}_{j,b}\left(w_{j,b}\right)-F_{j,b}\left(w_{j,b}\right)\right|=o_p\left(1\right)$$

In addition,

$$\begin{aligned}F_{j,b}\left(w_{j,b}\right) &= \Pr\left[W_{j,b}\leq w_{j,b}\right] \\ &= \Pr\left[W_j\cdot c_{jb}\leq w_j\cdot c_{jb}\right] \\ &= F_j\left(w_j\right)\end{aligned}$$

Note that

$$\begin{aligned}\sup_{w_j\in\Theta}\left|\Gamma_j\left(w\right)-\widehat{\Gamma}_{j,b}\left(w\right)\right| \leq &\sup_{w_j\in\Theta}\left|\Gamma_j\left(w\right)-\widehat{\Gamma}_j\left(w\right)\right| \\ &+\sup_{w_j\in\Theta}\left|\widehat{\Gamma}_j\left(w\right)-\widehat{\Gamma}_{j,b}\left(w\right)\right|\end{aligned}$$

where $\widehat{\Gamma}_j\left(w\right)=\int...\int 1\left\{\sum_{j=1}^{N_1}W^{R}_j<w\right\}d\widehat{F}_1\left(W^{R}_1\right)\cdot d\widehat{F}_2\left(W^{R}_2\right)\cdot...\cdot d\widehat{F}_{N_1}\left(W^{R}_{N_1}\right)$. By the results above,

$$\sup_{w_j\in\Theta}\left|\Gamma_j\left(w\right)-\widehat{\Gamma}_j\left(w\right)\right|=o(1)$$

$$\sup_{w_j\in\Theta}\left|\widehat{\Gamma}_j\left(w\right)-\widehat{\Gamma}_{j,b}\left(w\right)\right|=o_p\left(1\right)$$

Now, we show that $\Pr\left[d^{*}_{1-\frac{\alpha}{2}}\leq\widehat{\alpha}-\alpha_0\leq d^{*}_{\frac{\alpha}{2}}\middle|\alpha_0\right]\to_p 1-\alpha$. As $N_0\to\infty$,

$$\widehat{\alpha}-\alpha_0\to\frac{1}{N_1}\sum_{j=1}^{N_1}\widehat{W}^{R}_j \text{ and } \widehat{\alpha}_b-\alpha_0=\frac{1}{N_1}\sum_{j=1}^{N_1}\widetilde{W}_{j,b}$$

Using the results above, we can show that

$$\begin{aligned}\Pr\left[d^*_{1-\frac{\alpha}{2}} \leq \widehat{\alpha} - \alpha_0 \leq d^*_{\frac{\alpha}{2}} \middle| \alpha_0\right] &= \Pr\left[d^*_{1-\frac{\alpha}{2}} \leq \widehat{\alpha}_b - \alpha_0 \leq d^*_{\frac{\alpha}{2}} \middle| \alpha_0\right] + o_p(1) \\ &= 1-\alpha\end{aligned}$$

■

The approach proposed to estimate $\widetilde{W}_{j,b}$ is unfeasible since we do not the variances of W_j's. Theorem 2 shows that if we have a consistent estimator of $\sqrt{\frac{Var\left[W_j^R\right]}{Var\left[W_{j,b}^R\right]}}$, we can construct $\widehat{\widetilde{W}}_{jb} = \widehat{W}^R_{jb} \cdot \sqrt{\frac{\widehat{Var\left[W_j^R\right]}}{\widehat{Var\left[W_{j,b}^R\right]}}}$, and use the approach proposed above.

Theorem 2 *Define $d^*_{1-\frac{\alpha}{2}}$ and $d^*_{\frac{\alpha}{2}}$ as the $(1-\frac{\alpha}{2})$th and $\frac{a}{2}$th quantile of the empirical distribution of $(\widehat{\alpha}_b - \alpha_0)$, $b = 1, ..., \mathcal{B}$. If for each j $\sqrt{\frac{\widehat{Var\left[W_j^R\right]}}{\widehat{Var\left[W_{j,b}^R\right]}}}$ is a consistent estimator for $\sqrt{\frac{Var\left[W_j^R\right]}{Var\left[W_{j,b}^R\right]}}$, under assumptions 1 and 2,*

$$\Pr\left[d^*_{1-\frac{\alpha}{2}} \leq \widehat{\alpha} - \alpha_0 \leq d^*_{\frac{\alpha}{2}} \middle| \alpha_0\right] \to_p 1-\alpha$$

Proof. Now, we do not know the variance of $\not{W}_j$.In this case, we define $\widehat{F}_j(\widehat{w}_j) = \frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}} 1\{\widehat{\widetilde{W}}_{j,b} < w_j\}$

$$\begin{aligned}\sup_{w_j\in\Theta}\left|\widehat{F}_j(w_j) - \Gamma_j(w)\right| &= \sup_{w_j\in\Theta}\left|\widehat{F}_j(\widehat{w}_j) - \widehat{F}_j(w_j) + \widehat{F}_j(w_j) - \widehat{F}^*_j(w_j) + \widehat{F}^*_j(w_j) - \Gamma_j(w)\right| \\ &\leq \sup_{w_j\in\Theta}\left|\widehat{F}_j(\widehat{w}_j) - \widehat{F}_j(w_j)\right| + \sup_{w_j\in\Theta}\left|\widehat{F}^*_j(\widehat{w}_j) - \widehat{F}^*_j(w_j)\right| + \sup_{w_j\in\Theta}\left|\widehat{F}^*_j(w_j) - \Gamma_j(w)\right|\end{aligned}$$

We show in the previous theorem that $\sup_{w_j\in\Theta}\left|\widehat{F}^*_j(\widehat{w}_j) - \widehat{F}^*_j(w_j)\right| = o(1)$ and $\sup_{w_j\in\Theta}\left|\widehat{F}^*_j(w_j) - \Gamma_j(w)\right| = o_p(1)$. We only need to work with the first term,

$$\begin{aligned}\sup_{w_j\in\Theta}\left|\widehat{F}_j(\widehat{w}_j) - \widehat{F}_j(w_j)\right| &= \sup_{w_j\in\Theta}\left|\frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}} 1\{W^R_{j,b} < w_j\cdot\widehat{c}_{jb}\} - \frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}} 1\{W^R_{j,b} < w_j\cdot c_{jb}\}\right| \\ &\leq \frac{1}{\mathcal{B}}\sum_{b=1}^{\mathcal{B}}\sup_{w_j\in\Theta}\left|1\{W^R_{j,b} < w_j\cdot\widehat{c}_{jb}\} - 1\{W^R_{j,b} < w_j\cdot c_{jb}\}\right| \\ &\to_p 0 \text{ since } \widehat{c}_{jb} \to_p c_{j_b}.\end{aligned}$$

■

We proposed a consistent estimator of $\sqrt{\frac{\widehat{Var\left[W_j^R\right]}}{\widehat{Var\left[W_{j,b}^R\right]}}}$ based on an ordinary least squares estimator. We estimate a linear regression that relates $\left(\widehat{W}^R_j\right)^2$ with $\frac{1}{M_j}$ and constant. We obtain $\widehat{A}$ as the least squares coefficient associated with the constant, and $\widehat{B}$ as the coefficient associated with $\frac{1}{M_j}$. We use A and B to construct a consistent estimator for the $Var[W_j^R]$,

$$\widehat{Var\left[W_j^R\right]} = \widehat{A} + \frac{\widehat{B}}{M_j}$$

We use these two estimator to estimate the ratio $\widehat{c}_{jb} \equiv \sqrt{\frac{\widehat{Var\left[W_j^R\right]}}{\widehat{Var\left[W_{j,b}^R\right]}}}$. Theorem 3 shows that is $\widehat{c}_{jb}$ is a consistent estimator for $\sqrt{\frac{Var[W_1]}{Var[W_j]}}$.

Theorem 3 *Under assumptions 1 and 2, $\widehat{c}_j$ is a consistent estimator for $c_{jb} = \sqrt{\frac{Var\left[W_{j,b}^R\right]}{Var\left[W_j^R\right]}}$.*

Proof. *By assumption 2,*

$$Var\left[W_j^R\right] = A + \frac{B}{M_j} \text{ and } \mathbb{E}\left[W_{jt}\right] = 0$$

So we can write

$$\mathbb{E}\left[\left(W_j^R\right)^2\right] = A + \frac{B}{M_j}$$

or

$$\left(W_j^R\right)^2 = A + \frac{B}{M_j} + \omega$$

where $\mathbb{E}\left[\omega\right] = 0$. In this case, we estimate A and B by ordinary least squares, we obtain consistent estimators as $N_0 \to \infty$. Since M_j does not vary with N_0, $\widehat{g}\left(M_j\right) \to_p g\left(M_j\right)$. ■